6/20/2014

Explicit Anomalous Cognition: A Review of the Best Evidence in Ganzfeld, Forced-choice, Remote Viewing and Dream Studies

Johann Baptista, Max Derakhshani and Patrizio Tressoldi

To appear in

The Parapsychology Handbook.

Etzel Cardeña, John Palmer, and David Marcusson-Clavertz Eds.

In its quest to demonstrate the reality of psi phenomena to the mainstream scientific community, parapsychology has rarely been afforded the luxury of dealing in averages. On the contrary, the field has had to develop a progressive ethos, a long and distinguished tradition of adopting the more rigorous and innovative techniques in science—sometimes even to the point of creating them. It is no coincidence, for example, that the first comprehensive meta-analysis in scientific history was performed by Rhine et al. on ESP card studies (Bösch, 2004), or that the first official policy of publishing null results was set out by the Parapsychological Association (Carter, 2010). Neither is it trivial that psi research has kept pace with associated mainstream and behavioral fields in terms of reproducibility—on a minimal budget (Baptista & Derakhshani, 2014); exceeded their criteria with respect to masking practices (Watt & Nagtegaal, 2004); and exceeded them again in terms of reporting negative results (Mosseau, 2003)[1] . It should indeed be this way, for not only has parapsychology been subject to (and strengthened by) an intensified back-and-forth between proponents and skeptics along its history, but the claims it propounds have always demanded high standards of evidence. This is a fact recognized by sides of the psi debate.

Our contribution to the Handbook is an attempt to further this point of agreement. We propose changes that will add constructively to the face validity of parapsychological claims by combating potential flaws before they occur—changes informed by our review of four key, conceptually simple domains of experimentation: ganzfeld, forced-choice ESP, Remote Viewing, and dream ESP. Subsequent to our review of these studies, we provide further suggestions on the statistical treatment of meta-analysis, applicable to ESP research generally.

Although the evidence for explicit anomalous cognition (EAC) so far produced in parapsychology has been resistant to criticism and demonstrated compelling trends (Storm, Tressoldi, & DiRiso, 2010; Tressoldi, 2011; Baptista & Derakhshani, 2014), in the face of our analysis we find the most prudent course of action is to agree with our critics that the evidence can be—and ought to be—improved, if for no other reason than that improvement is inherent to the parapsychological enterprise. Improvement consists in obviating potential flaws before they have the chance to occur (whether or not they do) and/or in making the EAC effect more robust, generally (so as to *a priori* rule out explanations based on flaws).

Ganzfeld

We review the ganzfeld studies first by summarizing their history from the early 1970s until the early 2000s. Then we focus on the most contemporary evidence available, from the Storm, Tressoldi, and DiRisio (STDR; 2010) database. Afterwards, we present several new meta-analytic findings from Baptista and Derakhshani (2014) relevant to the current state of the research—as well as to other altered states EAC studies.

Early History

From 1974 to 1981, 42 ganzfeld studies were conducted by 47 different investigators, analyzed by both Hyman (1985) and Honorton (1985). Hyman provided a skeptical appraisal, identified methodological weaknesses, and performed a factor analysis that found evidence of a significant

[1] For a comprehensive overview of various benchmarks of good science, as well as how parapsychology fares on them, we refer the interested reader to the cited papers, with emphasis on Mosseau (2003).

correlation between specified flaw indices and ganzfeld rates of success. In response, Honorton criticized Hyman's flaw categorizations and asked psychometrician David Saunders to comment on his method, to which Saunders (1985) stated that factor analysis was inappropriate given the small sample size of Hyman's database. Honorton also attempted to counter Hyman's multiple-analysis criticism by meta-analyzing only the 28 ganzfeld studies in the original database of 42 that used direct hits as their measure of success. He found an unweighted Stouffer's $Z = 6.6$, $p = 10^{-9}$, where 43% of the studies were significant at the $p < .05$ level, with a hit rate (HR) of 36.87% for the 26 studies of four-choice design. The debate ended with the publication of the Joint Communiqué (Hyman & Honorton, 1986), in which the authors agreed that there was an overall effect in the database but differed on the extent to which it constituted evidence for psi. Stricter methodological guidelines were proposed by both authors, and a consensus was reached that "the final verdict will await the outcome of future experiments conducted by a broader range of investigators and according to more stringent standards." *p*. 351.

Eight years later, Bem and Honorton (1994) published a meta-analysis of 10 automated ganzfeld studies conducted in Honorton's Psychophysical Research Laboratories (PRL) designed to meet these more stringent standards, comprising 329 sessions in total, with a $Z = 2.89$, $p = .002$, and an hit rate of 32.2%. This successful effort made inroads towards the aims of the Joint Communiqué and was promoted by the authors as evidence that the ganzfeld psi effect was both robust and reproducible.

However, a later meta-analysis by Milton and Wiseman (1999) conducted on all the studies that came after the PRL (from 1987 to the later cutoff of 1997) did not yield similar findings. Comprising 30 studies, it reported a null result: $Z = 0.70$, $ES = 0.013$, $p = .24$, hit rate = 27.55%. Milton and Wiseman surmised from this that the PRL had not been replicated, and that the ganzfeld paradigm did not provide evidence for psychic functioning—a conclusion that sparked an involved debate in the parapsychology community (Schmeidler & Edge, 1999), and was frequently mentioned in skeptical circles as evidence of the irreproducibility of psi.

Although the appropriateness of the statistical test in Milton and Wiseman (1999) has been questioned (Carter, 2010; Baptista & Derakhshani, 2014), its identification of a significantly reduced overall effect size for the ganzfeld database has not. The hit rate drop from 32.2% (PRL) to 27.6% (Milton & Wiseman, 1999) required an explanation.

Two years afterwards, a potential one was supplied by Bem, Palmer, and Broughton (2001). Bem et al. presented a meta-analysis of their own, arguing that the reason for Milton and Wiseman's (MW's) effect size plunge was that many studies subsequent to the PRL studies had employed novel, nonstandard procedures that were at greater risk for failure (such as Willin, 1996, which used musical targets). Their meta-analysis added 10 new studies published after the MW cut-off and included a 7-point standardness scale for each study, where ranks were awarded by masked raters. For this database of 40 studies, Bem et al. (2001) found a hit rate of 30.1%, an $ES = 0.051$, and a Stouffer's $Z = 2.59$, $p = .0048$. Additionally, their hypothesis was supported by the fact that those studies that ranked above the midpoint (4.0) for standardness yielded significant results at a hit rate of 31.2% (1278 trials, 29 studies, exact binomial $p = .0002$) and those that fell below the midpoint supplied a nonsignificant hit rate of 24%—and the difference was significant ($U = 190.5$; $p = .020$). But their analysis possessed one weakness: with the addition of the 10 new studies, one of which was the very successful Dalton (1997) study with artistic subjects—hit rate of 47% and 128 trials—their new hit rate of 30.1% was already independently highly significant.

Additionally, as Baptista and Derakhshani (2014) argued, since almost 100% of PRL subjects had at least one of the psi conducive traits used by Storm et al. (2010) in their criteria for selected subjects[2] (88% personal psi experience; 99.2% above midpoint on psi belief; 80% training in meditation; Honorton, 1990), it is appropriate to ask what the hit rate was for subjects who possessed these traits in the Milton and Wiseman database. It turns out to be between 30.88% and 34.2% (see next section). So Bem, Palmer, and Broughton's result should be taken with a grain of salt; nevertheless, that the hit rate for their studies rose to 33% (974 trials, 21 studies) for experiments that ranked 6 or above on the similarity scale—almost exactly PRL's own hit rate—shows that study standardness at least captured something that correlated positively and reliably with experimental results.

Aside from Bem, Palmer, and Broughton's (2001) comparative meta-analysis, one other meta-analysis was published in the same year by Storm and Ertel (2001) which included all 79 reported studies from 1982 to 1997, combining the 10 PRL studies in Bem and Honorton (1994) with Honorton's (1985) and Milton and Wiseman's (1999) databases: $ES = 0.138$, Stouffer's $Z = 5.66$, $p = 7.78 \times 10^{-9}$. They criticized Milton and Wiseman's method of analyzing only the post-PRL studies and concluded that, despite their dip in effect sizes, the ganzfeld experiment remained a viable method for eliciting psi and would benefit from further replication.

A Reanalysis of the Milton and Wiseman Database

As was commented above, it is appropriate to ask what percentage of participants in the Milton and Wiseman database had "selected" characteristics, and what their success rate was. The answer to this question is not so straightforward. In procuring an approximate one, it is well to make a distinction between "selected subjects"—participants explicitly selected on the basis of Storm et al.'s criteria—and "happenstance" subjects who fortuitously possessed at least one psi-conducive trait. This distinction is important because the PRL studies themselves did not make *formal* efforts to recruit by trait; nevertheless, their use of personal referrals, experimenter-suggested participants, recruitment at PRL presentations, etc, *a priori* suggests a pool of subjects that is more highly motivated and psi believing than—for instance—a pool resulting from advertisements to undergraduate psychology students. Their census confirms this intuition. Of course, any parapsychology study invariably includes at least some subjects who possess such traits, so it might seem that to delineate a full inclusion criteria is an impossible task.

One approach to solve this problem is to consider the PRL studies effectively *selected*, given the undeniably high occurrence of psi-conducive traits in their subjects. This choice is desirable because of its conservativeness; to ascertain replication, only studies from the Milton and Wiseman database which observe the criteria of psi-conduciveness to a *greater* degree than the PRL (i.e. 100% of at least one trait must be present) are included, and there is no ambiguity in that inclusion. When considered this way, there are only three studies from Milton and Wiseman (1999) that fulfill that criterion (Morris, Cunningham, McAlpine & Taylor, 1993; McDonough et al., 1994; Morris et

[2] From Storm et al. (2010): "... participants who had previous experience in *ES*P experiments, or were psi believers, or had special psi training, or were longtime practitioners of meditation or relaxation exercises, etc...". Given Storm et al.'s use of creative participants as selected, we also include those. This is just as well considering the use of creative subjects in the PRL studies 104 and 105.

al., 1995), with a total sample size of 149 and a hit rate of 34.2%, significant at $p = .007$. This is suggestive of replication, but with only three studies clearly the matter is not settled (it is nevertheless useful to point out that these 3 studies make up only 10% of the Milton and Wiseman 1999 database, whereas in the more successful Storm et al. 2010 database the proportion of selected studies is 47%). To augment the number of studies and of trials we may attempt to add happenstance subjects, a strategy complicated by the fact that in most studies where records were given on participant traits they were breakdowns of individual traits, and substantial overlap was present.

In order to overcome this, we have to make a decision about which trait to include that is objective, and a reasonable criterion for this decision is to include only the data from the trait with the largest number of participants. If this is done, we can tentatively add trials from three more studies[3], for a combined total of 513 trials and a hit rate of 30.60%, significant at—again—$p = .007$, and non-significantly different from the 32.2% hit rate of the PRL; Fisher's exact $p = .65$, two tailed. Together, these observations make a good case that the PRL results were replicated in the Milton and Wiseman database when similarity of populations alone is considered, without necessary reference to experimental standardness (as reported in Bem, Palmer, and Broughton, 2001).

The Storm-Tressoldi-DiRisio Database

Almost a decade after the 2001 Milton and Wiseman study, the ganzfeld was once again in need of a systematic review. To accomplish this, Storm et al. (2010) meta-analyzed 30 studies that had been conducted from 1997 to 2008, contributed by 36 different investigators, comprising 1,648 trials, within a larger meta-analysis of free-response ESP studies conducted during that period. They selected for inclusion only those ganzfeld studies that had more than two participants, used a random number generator or a random number table for target selection, and provided enough information to calculate direct hits. The result was a mean $Z = 1.16$, Stouffer's $Z = 6.34$, $ES = 0.152$, and $p = 1.15 \times 10^{-10}$, for their 30-study heterogenous database; and a mean $Z = 1.02$, Stouffer's $Z = 5.48$, $ES = 0.142$, and $p = 2.13 \times 10^{-8}$, for their 29-study homogenous database (in which the authors removed outliers using the stem-and-leaf plot method). Storm et al. used this latter, homogenized database to perform several important analyses: (a) a comparison of the effectiveness of competing experimental conditions (i.e. ganzfeld, non-ganzfeld noise reduction, and standard-free response), (b) an assessment of the performance of selected vs. unselected participants, (c) a test of experimenter effects, and (d)-(e) file-drawer assessments.

For analysis (a), Storm et al. (2010) hypothesized that the mean *ES* values of their three study categories would be arrayed in sequential order, ganzfeld having the highest *ES* (.142), followed by non-ganzfeld noise reduction (.110), and then standard free-response (.029). These results suggest that the ganzfeld procedure is still the most well-developed of the free-response categories of psi studies. They also are consistent with the hypothesis that sensory isolation elicits increased psychic functioning, although only the difference between the ganzfeld and the standard free response studies was significant.

[3] 68 mental disciplines subjects from Bierman (1993), with 21 hits; all 151 subjects and 40 hits in Broughton & Alexander (1997) because of similar numbers to PRL database (91.4% psi believing, 74.2% practice of metnal discipline); and 145 psi-believing subjects from Kanthamani & Broughton (1994) with 45 hits.

Analysis (b), a univariate ANOVA comparison of the performance of selected participants vs. unselected participants across all three types of studies, yielded no significant difference between *ES* values for these subgroups; *p* = .14, two-tailed. However, since Storm et al. wrote in their paper that their hypothesis was of better performance in the selected group, an argument can be made that their two-tailed test should have been replaced by a one-tailed test, whose results would have approached significance; *p* = .07, one-tailed, suggestive of a difference. The ANOVA analysis as reported did reveal a significant category effect; upon further examination, Storm et al. (2010) found that was due entirely to the ganzfeld condition, in which selected participants (*ES* = 0.26) had outperformed unselected participants (*ES* = 0.05) by half an order of magnitude—and the difference was statistically significant via a separate t-test; *t*(27) = 3.44, *p* = .002. This evidence tentatively supports previous findings in parapsychology that participant selection is a moderator variable of psi performance, generally.

For analysis (c) of experimenter effects, Storm et al. (2010) divided the ganzfeld and non-ganzfeld noise reduction studies into seven mutually exclusive experimenter/laboratory groups, with at least two studies in each. Those groups were: “Morris”, “Parker”, “Parra”, “Roe”, “Roney-Dougal”, “Tressoldi”, and “Wezelman”. No significant differences between *ES* values were observed for these groups, suggesting that experimenters do not have a major effect on study outcomes in the ganzfeld; *p* = .315.

Finally, for analyses (d)-(e), Storm and his colleagues estimated the severity required for file-drawer effects to nullify their database. Analysis (d) was a standard Rosenthal Fail-Safe calculation, which revealed that approximately 293 excluded studies averaging a *z*-score of zero would be required to bring their significant finding to a chance result. Analysis (e), however, applied a more conservative calculation by Darlington and Hayes (2000) that allowed for many of the potentially excluded studies to have “highly negative” *z*-scores. This analysis determined that at least 95 missing studies would be needed to bring Storm et al.'s meta-analysis to nonsignificance—86 of which could have negative *z*-scores. Selective reporting, therefore, is unlikely to account for the ganzfeld findings.

Storm et al. (2010) also combined their 29-study database with Storm and Ertel’s (2001) earlier 79-study database, comprising data from 1974 to 2001. Having found no significant difference in *z*-scores between these two sets, they formed a heterogenous aggregate of 108 studies, with a mean $Z = .80$, Stouffer’s $Z = 8.31$, $ES = .142$, and $p = 8.31 \times 10^{-10}$. Then they homogenized their database by removing six outliers identified through SPSS stem-and-leaf and box-and-whiskers plots, bringing the number of studies down to 102; mean $Z = .81$, Stouffer’s $Z = 8.13$, $ES = .135$, $p < 10^{-16}$.

It was the larger, heterogenous database that Storm et al. (2010) used to test for decline effects across the ganzfeld domain (with studies from 1974 to 2008). By applying linear regression, they concluded that although a slight negative and significant correlation was present (Figure 1) between study year and study *ES*, *r*(106) = -.21, *p* = .049, the data were better fitted to a significant quadratic polynomial curve (Figure 2) of the form *ES* = 0.0009 YEAR2 + 3.4375 YEAR + 3.4272 ; $R^2 = .12$, *p* = .004; this indicated a nontrivial rebound effect in later years. Removing four outliers made the linear correlation nonsignificant; *p* = .126; although still negative, *r*(102) = -.15. The rebound effect observed in the non-linear model can be explained as the rise in effect size from the

Milton and Wiseman (1999) database to the Storm et al. (2010) database, itself independently statistically significant with a linear regression; $r = .27$, $p = .03$. Together, these findings suggest that the ganzfeld has not been strongly impacted by decline and remains viable as a method for experimenters to pursue in the future.

The Baptista-Derakhshani Analyses

Baptista and Derakhshani (2014) explored the validity of skeptical hypotheses for the Storm, Tressoldi, and DiRisio post-Milton-Wiseman database by analyzing the relations between important study variables and attempted to obtain empirical measures for the extent of selective reporting in the ganzfeld.

Exploring Relations Between Study Quality, Effect Size, and Year

The motivations for these analyses were to locate possible decline effects in the most recent database from 1997 to 2008 from Storm et al., to observe whether study quality had improved or declined, and to detect whether a negative correlation existed between study quality and effect size. The studies analyzed here form the most recent available database of ganzfeld studies; Baptista and Derakshani considered it important to analyze separately since the results of this database are the most likely to be informative for current researchers. The study quality ratings they used are those given by Storm et al.; we append the method and criteria of their ratings to the end of this section for convenient reference.

For the analyses, Baptista and Derakhshani (2014) first plotted study *ES*s against study publication year across the whole database and found no decline in *ES*s ($r = .00$). In contrast, for their *ES* vs. quality rating comparison, there was a positive and significant correlation, $r(28) = .37$, $p = .045$. That is, studies rated with higher methodological quality produced larger *ES*s than lower quality studies, and this trend was statistically significant.

Derakhshani (2014), however, found highly significant heterogeneity ($\chi^2 = 56.64$, $p = .0016$) for these studies, and discovered that blocking them according to whether they used selected or unselected participants produced two safely homogeneous sets. The selected participants subgroup included 14 studies of four-choice design, with a 40.1% overall hit rate across 748 trials; the unselected participants subgroup composed 15 studies of four-choice design, with a 27.3% overall hit rate across 886 trials. The difference in hit rates was extremely significant (Fisher's exact, $p < .0001$). Despite this, the mean quality rating of studies with selected participants, weighted by sample size, was not lower than the mean quality rating for unselected studies; it was higher (q = .84 and q = .79 respectively, where q = 1.00 is the highest possible rating)—but not significantly.

With such a heterogeneous database, correlations can be confounded; accordingly, Baptista and Derakhshani repeated their analyses on the two subgroups of selected and unselected participants. They found a small negative correlation between *ES* and study year for studies with selected participants, but it was not significant; $r(12) = -.30$, $p = .29$. They also found a positive, nonsignificant correlation; $r(12) = .27$, $p = .37$; between study quality and study *ES*, and a positive and nonsignificant correlation; $r(12) = .26$, $p = .37$; between study quality ratings and study year. Under the skeptical prediction, one would expect *ES*s to decrease across years as quality ratings increased, as observed in these analyses; however, a positive correlation between study quality and study *ES*, as was also observed, would not have been predicted. Given that these results are conflicting, and that none of the correlations were significant, Baptista and Derakhshani concluded that no reliable conclusions could be drawn about which correlations were real and which were spurious. More selected participant studies would be needed for this to be determined—although it could well be that both results are real, and confounding variables explain the discrepancy.

The unselected participants subgroup, on the other hand, had more striking results. For *ES* vs. year, Baptista and Derakhsani found a highly significant positive correlation, *r*(13) = .65, *p* = .007, two-tailed. For quality ratings vs. *ES* , they found a positive but nonsignificant correlation, *r*(13) = .40, *p* = .13, two-tailed. And for quality ratings vs. year, they found an extremely significant positive correlation, *r*(13) =.86, *p* < .0001, two-tailed. Baptista and Derakhshani were unable to provide reasons why these correlations were so much more significant than those of the selected participants subgroup, but the question surely merits further research.

To summarize, there are no reliable decline effects in either the entire Storm et al. (2010) post-Milton-Wiseman ganzfeld database or subgroups of selected and unselected participants in these databases. However, positive and significant correlations exist between quality and *ES* for both the post-Milton-Wiseman set as a whole and the unselected participants subset. We can therefore say with high confidence that the skeptical hypotheses involving declines in effect size and inverse relationships between quality and *ES*, examined by Baptista and Derakhshani (2014), are not supported by Storm et al.'s database.

Relationships Between *z*-scores, *ES*, √*N*, and *N*

Another important skeptical hypothesis for parapsychology generally was proposed by Kennedy (2003), namely that *ES* would drop for studies with larger sample sizes, and that there would be a negative correlation between study *z*-scores and the square root of study sample sizes. Both of these alleged effects suggest that psi slowly disappears over the course of long experiments, or—in the case of the latter—that publication bias is present. Derakhshani (2014) tested this claim on all three ganzfeld databases: pre-PRL, PRL, and post-PRL.

Looking at the 28 studies in Honorton's 1985 ganzfeld database, Derakhshani found that there was actually a slight non-significant incline effect; *r* = .13, *p* = .26, one-tailed; between study *z*-scores and study sample sizes, and a slight non-significant decline effect for study *ES*s and study sample sizes; *r* = -.14, *p* = .47, two-tailed. In other words, there was no clear decline or incline effect in the Honorton database. See Tables 1 and 2 for these results, appended to the end of this section.

Examining the 10 PRL studies, there was a significant decline effect in terms of study *ES*s vs. sample sizes. However, Bem and Honorton (1994) showed that this could be attributed to two particularly small studies that used highly selected participants and were predicted to do significantly better than participants in previous studies (the 20 session Juilliard study, and the 7 session Study 201). Consequently, the decline effect in their database appears to be artifactual.

In the 60 post-PRL studies, Derakhshani (2014) found that there was a moderate; *r* = .39; and highly significant incline effect; *p* = .001, one-tailed; between study *z*-scores and the square root of study sample sizes, and a small incline effect between study *ES*s and study sample sizes; *r* =.22, *p* = .09, two-tailed. See again Tables 1 and 2 for these results.

For the most recent post-PRL database subset of 30 studies (or, alternatively, the post-MW studies), Derakhshani (2014) found an even stronger and highly significant incline effect between study *z*-scores and the square root of study sample sizes; *r* =.47, *p* = .004, one-tailed; and a small incline effect between study *ES*s and study sample sizes; *r* = .20, *p* = .29, two-tailed. See again Tables 1 and 2 for these results. In addition, by grouping the studies that used only selected participants (14 studies), Derakhshani (2014) found a much stronger and extremely significant incline effect between study *z*-scores and the square root of study sample sizes; *r* = .83, *p* = .0001, one-tailed; and a large and nearly marginally significant incline effect between study *ES*s and study sample sizes; *r* = 0.41, *p* = .14, two-tailed. There are also incline effects for the studies using only unselected participants, but they are significantly smaller than for studies using only selected participants. See Tables 6 and 7 for these results.

Finally, examining the Storm et al. database of 108 studies (which includes all the ganzfeld databases we have considered), Derakhshani (2014) found an overall significant incline effect between study *z*-scores and the square root of study sample sizes; *r* = .20, *p* = .02, one-tailed, as well as a slight incline effect between study *ES*s and study sample sizes; *r* = .05, *p* = .65, two-tailed. See again Tables 1 and 2 for these results.

In sum, (a) the pre-communiqué studies (i.e. the studies in Honorton's database) show no clear incline or decline effect; (b) the PRL studies show an artifactual decline effect; (c) post-PRL studies consistently show moderate to large incline effects in all cases; (d) the Storm et. al (2010) homogeneous database of 102 studies shows a small but significant incline effect for the square root of study sample size vs. study *z*-score, and study *ES* vs. study year. For readers who wonder whether heterogeneity may have been present in the databases discussed in (a) – (c), and if removal of this heterogeneity may affect the conclusions, we present Derakhshani's (2014) analyses for the homogeneous versions of those databases in Tables 3-5, which show that we reach the same conclusions.

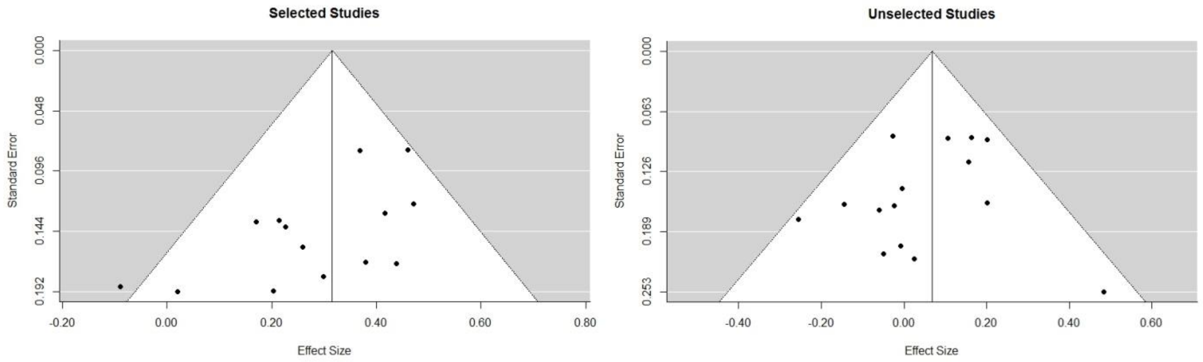
The findings in (c) are arguably more important and reliable for making generalizations about the ganzfeld ESP effect, as the post-communiqué studies are significantly better in methodological quality than the pre-communiqué studies, and there are more than twice as many post-communiqué studies than pre-communiqué studies—the former studies having much larger sample sizes on average.

Based on all these considerations, neither the *z* vs. $\sqrt{N}$ decline effect nor the *ES* vs. *N* decline effect have been shown to occur in the ganzfeld database, whereas significant incline effects have been demonstrated to occur in the post-PRL database. This is consistent with the hypothesis that the ganzfeld effect behaves lawfully, in accordance with the assumptions of power analysis.

Gauging the File-Drawer

For the task of estimating selective reporting, Baptista and Derakhshani (2014) found the systematic review of Watt (2006) helpful; it surveyed all of the parapsychology undergraduate projects undertaken and supervised at the Koestler Parapsychology Unit (KPU) in Edinburgh, Scotland, between 1987 and 2007[4] . Because of Watt's survey, the KPU ganzfeld pool is a good example of a dataset that can be reasonably inferred to possess no excluded studies. Considering the five studies provided[5]; a total of 195 trials, 66 hits, and a hit rate of 33.8%; they obtained an exact binomial *p* = .004, one-tailed. The 10-study PRL database, too, is known to have no selective reporting[6] , with a hit rate of 32.2%, 329 trials, 106 hits, and a binomial probability of *p* = .002. Given that these hit rates are not significantly different from each other, Baptista and Derakhshani merged the two datasets, forming one overall 15 study pool with no file drawer, 524 trials, 172 hits, a hit rate of 32.8%, and a binomial probability of $p = 5.91 \times 10^{-8}$. They pointed out that this composite hit rate (32.8%) was close to that of the remaining 90 studies in Storm et al.'s (2010) total heterogeneous database—removing these 15 studies as well as 3 not of four-choice design—for a composite hit rate of 31.8%, across 3,516 trials. This convergence of results from three analyzed study pools (the KPU, the PRL, and rest of the ganzfeld) suggests that if there is a contribution from selective reporting to the overall hit rate, it is likely to be minimal.

The prevalence of the file-drawer effect can also be estimated through the standard funnel plot, displaying study precision (standard error) against study *ES*; as the precision of the study increases, the variation in the effect sizes decreases, leading to a symmetrical inverted funnel shape. Missing studies can then be identified through plot asymmetry, usually as a lack of small studies with small effect sizes, on the bottom left side of the plot. However, inferences from asymmetry become less reliable with more heterogeneity (Sterne et al., 2011), so we plotted the two homogenous groups of selected and unselected participant studies in the Storm et al. (2010) database, to counter this problem:



In neither plot is there evidence that small studies with small effect sizes have been excluded, but since the number of studies is not very large, a perfect funnel should not be expected.

[4] Data is taken from the paper that was updated and presented at the 2007 PA convention.

[5] Colyer and Morris, 2001; Morris, Cunningham, McAlpine, and Taylor, 1993; Morris, Summers, and Yim, 2003; Symmons and Morris, 1997. It should be noted that Morris,Cunningham, McAlpine, and Taylor (1993) comprised two projects, hence why only four studies are listed.

[6] Bem and Honorton (1994) explicitly state that "the eleven studies just described comprise all sessions conducted during the 6.5 years of the program. There is no file-drawer of unreported sessions"(*p*. 10). Additionally, Honorton (1985) also states, "Except for two pilot studies, the number of participants and trials was specified in advance for each series. The pilot or formal status of each series was similarly specified in advance and recorded on disk before beginning the series. We have reported all trials, including pilot and ongoing series, using the digital autoganzfeld system. Thus, there is no 'file-drawer' problem in this database."(*p*.133)

A final note should be added to the discussion of publication bias, concerning the validity of Rosenthal's fail-safe calculation. Specifically, the calculation has been often criticized for its assumption of a neutral file drawer (i.e., mean $Z = 0$), with the charge that such a proposition is both practically and mathematically unjustifiable for a null distribution. Scargle (2010), for example, demonstrated that Rosenthal's assumption should be corrected to mean $Z = -0.1085$, not 0, given a 95% cutoff for lower z-scores (i.e., those nonsignificant according to the $p > .05$ criterion). This is indeed true for a null distribution, but as Rosenthal and Harris (1988) point out, the z-scores of nonsignificant studies in most fields are typically pulled strongly towards the mean z of the studies meta-analyzed, meaning that if studies are rejected for inclusion at the cutoff of $p > .05$ (according to the classic file-drawer hypothesis), the file-drawer itself is still likely to be positively skewed if even a moderate effect with moderate power is present. This would make Rosenthal's mean $Z = 0$ assumption a conservative estimate in most areas, including the ganzfeld[7] .

Quality Criteria and Tables

Below we append Storm et al's (2010) methodology for rating the quality of their studies as well as several tables pertaining to the aforementioned analyses, (a)-(c).

Storm et al.'s (2010) quality ratings were made by two judges (graduate students of Tressoldi) who saw only the method sections of each study article they assessed, from which all identifiers had been deleted (such as article titles, authors' hypotheses, and references to results of other experiments in the article). The following were their criteria:

1. Appropriate randomization (using electronic selection or random tables).
2. Random target positioning during judgment (i.e., target was randomly placed in the presentation with decoys).
3. Masked response transcription, or impossibility of knowing the target in advance.
4. Number of trials pre-planned.
5. Sensory shielding from sender (agent) and receiver (perceiver).
6. Target independently checked by a second judge.
7. Experimenters masked to target identity.

The two judges answered "yes" or "no" to each of the criteria. The study quality ratings were then defined as the ratio of points awarded with respect to the items applicable (minimum rating was 1/7 = 0.14; maximum rating was 7/7 = 1.00), and the quality ratings of each judge were averaged together.

Storm et al. (2010) reported a Cronbach's alpha for the two judge's ratings of .79, indicating high interrater reliability. Their criteria for study quality and their method of determining quality scores seem reasonable to us, and we cannot see any major flaws that might nullify our findings for either the unselected or selected participant subgroups.

[7] See Baptista & Derakhshani (2014) for a more detailed look at the applicability of Rosenthal's statistic to the ganzfeld.

Table 1: √*N* vs *z*

Database	*R*	*Df*	*t*	*p (one-tailed)*
Honorton '85[1]	.13	26	.67	.26
M & W 1999[2]	.14	28	.74	.23
STDR ('97-'08)[3]	.47	28	2.82	4.4×10^{-3}
Post-PRL[4]	.39	58	3.24	$1 \text{ x } 10^{-3}$
STDR ('74-'08)[5]	.20	106	2.05	.022

[1]Honorton (1985), [2]Milton and Wiseman (1999), [3]Storm et al. (2010) studies from 1997-2008, [4]Composite of Milton and Wiseman (1999) and most recent Storm et al. (2010) studies, [5]Storm et al. (2010) studies from 1974 to 2008.

Table 2: *N* vs. *ES*

Database	*R*	*Df*	*t*	*p (one-tailed)*
Honorton '85	-.14	26	-0.73	.47
M & W 1999	.11	28	0.61	.55
STDR ('97-'08)	.20	28	1.08	.29
Post-PRL	.22	58	1.71	.091
STDR ('74-'08)	.045	106	0.46	.65

For the second category of analyses, Tables 1 and 2 show the results of the linear regression for the square root of study sample size vs. *z*-score (√*N* vs. *z*), and study sample size vs. effect size (*N* vs. *ES*); with Honorton's (1985) database, the Milton and Wiseman (1999) database, the Storm, Tressoldi, and DiRisio (STDR; 2010) database of 30 studies from 1997–2008, the post-PRL database, and the STDR database of 108 studies from 1974–2008; all without outliers removed.

Derakhshani also assessed whether each database had a heterogeneous or homogeneous *ES* distribution relative to the sample size weighted mean *ES* (Hedges, 1981; Honorton et al., 1990; Rosenthal, 1986). The χ^2 formula he used is:

$$\chi^2(k+1) = \sum_i^k N_i(r_i - r)^2$$

where k is the number of studies, N_i is the sample size of the ith study, r_i is the *ES* (i.e. correlation coefficient) of the ith study, and the weighted mean *ES* is

$$r = \frac{\sum_i^k N_i r_i}{N}$$

Table 3: χ^2 Test

Database	*Df*	χ^2	*p (one-tailed)*
Honorton '85	27	-50.6	$3.8 \text{ x } 10^{-3}$
M & W 1999	29	041.3	.065
STDR ('97-'08)	29	156.6	$1.6 \text{ x } 10^{-3}$
Post-PRL	58	1112.6	$< 1 \text{ x } 10^{-4}$
STDR ('74-'08)	107	0203.7	$< 1 \text{ x } 10^{-4}$

Table 4: √*N* vs. *z*

Database	*R*	*Df*	*t*	*p (one-tailed)*
Honorton ’85	-.00	23	-.00	.50
STDR (’97-’08)	.40	26	12.18	.019
Post-PRL	.24	54	11.80	.039
STDR (’74-’08)	.11	96	02.05	.14

Table 5: *N* vs *ES*

Database	*R*	*Df*	*t*	*p (two-tailed)*
Honorton ’85	-.31	23	-1.56	.13
STDR (’97-’08)	.16	26	11.18	.24
Post-PRL	.11	54	1.81	.42
STDR (’74-’08)	.07	96	0.75	.45

Table 3 shows the results of these chi square tests, and Tables 4 and 5[8] show the regression analysis results for the outlier-removed homogeneous databases.

Along with removing outliers, another common way of dealing with heterogeneity is to identify moderator variables, and block the studies using those moderator variables. Derakhshani (2014) found that selected and unselected participants served as moderator variables for the heterogeneity in STDR’s ’97–‘08 database. More specifically, by blocking the studies into two subgroups—one made of selected participants (14 studies) and the other made of unselected participants (16 studies)—each subgroup became homogeneous; the overall hit rates being 40.1% for selected and 27.3% for unselected (for studies of four-choice design). The difference between these hit rates is extremely significant (Fisher’s exact $p < .0001$, two-tailed). Tables 6 and 7 below show the regression analysis results for each subgroup.

Studies with selected groups produce a very strong correlation between the square root of sample size and *z*-score, and nearly as strong a correlation between sample size and *ES*, both of which are more than double the respective correlations for the studies with unselected groups (which themselves are also impressively large in magnitude).

Table 6. Studies with Selected Groups

Y vs X	*R*	*Df*	*t*	*P*
√*N* vs *z*	.83	12	5.13	.0001 (one-tailed)
N vs *ES*	.41	12	1.58	.24 (two-tailed)

Table 7: Unselected Studies

Y vs X	*R*	*Df*	*t*	*P*
√*N* vs *z*	.41	14	1.69	.06 (one-tailed)
N vs *ES*	.18	14	0.67	.51 (two-tailed)

[8] The studies in these Tables only include the studies of 4-choice design. This means some studies (3 out of all the studies) that were not of 4-choice design were excluded. However, some of these studies were already outliers, and those that were not made a negligible difference to the results of the regression analyses when removed.

Future Directions

On the basis of the above correlations for $\sqrt{N}$ vs. z and a predictive power model utilizing existing meta-analyses of ganzfeld studies, Derakhshani (2014) predicted that it should be possible to boost the replication rates of future ganzfeld studies from ~25% to as high as 80%. By following several prescriptions, most important among them being the exclusive use of selected participants in all—or as many possible—future ganzfeld studies, he suggests that this could be done while keeping the mean sample size of ganzfeld studies effectively the same. For example, if the selected participant overall HR of 40.1% is the expected hit rate in future studies with selected groups, an investigator would require only 56 trials (about the mean size of the post-PRL studies) to achieve a statistical power of 80% at the 5% level. An important caveat is necessary here: this 40.1% overall hit rate for the Storm et al. (2010) database is very different from the hit rates of selected participants in past databases—approximately 32.2% for the effectively selected PRL and 34% for the selected studies in Milton and Wiseman (1999). This discrepancy may have several causes, and a satisfactory explanation of it would have to involve a comprehensive review of the differences between selected participants and experimental conditions across databases, which we do not provide here. Nevertheless, we can investigate the characteristics of selected participants who provide the largest effects, and we can stipulate that they may have been capitalized on in recent years to create the increase in the selected participant hit rate. This is merely a hypothesis in need of confirmation, but its validity does not influence the utility of finding strong-scorers.

Here, we do have some data to suggest which subjects are best. To begin with—as Baptista and Derakhshani (2014) report—in the PRL database; the Broughton, Kanthamani, and Khilji (1989) database; and the Kanthamani and Broughton (1994) database, the hit rates of subjects possessing at least one trait vs. three pre-specified traits (previous psi experience, a feeling-perception typology on the Myers-Briggs Personality Inventory, and practice of a mental discipline) were recorded, at 31 and 42% respectively, across all the independent databases. These facts support the intuitive idea that selecting subjects with multiple traits is superior to selecting for only one. Indeed, given that this combination of pre-specified traits—called by Honorton (1992) the "three-predictor model"—has already been tested prospectively with success, in two separate cases, we think it is very promising for use in future ganzfeld databases. Populations of creative and/or artistic subjects have fared equally well; the six studies that used them produced a combined hit rate of 41% in 367 trials (Derakhshani, 2014).

In sum, whenever possible, it is wise to make exclusive use of selected individuals. For a study with either the three-predictor model hit rate or the creative subjects hit rate, only between 44 and 50 trials are required for 80% power. Investigators should strive to use participants who are artists, musicians, twins, those who are biologically-related, emotionally close, have prior psi experience, mental discipline practice, prior psi training, belief in psi, and/or other critical characteristics.

Given the considerations reported above, our recommendations for future ganzfeld researchers are as follows:

1. Preplan the number of trials in a study on the basis of a power analysis to achieve at least 80% power. It is recommended that effect size estimates be conservative.

2. Keep ganzfeld trials close to methodologically standard (based on the results of Bem, Palmer, and Broughton, 2001)
3. Pre-register studies in any one of the registries available to parapsychology researchers (Open Science Framework, 2014; Koestler Parapsychology Unit (KPU) Registry, 2014).
4. Log as much information about participants and methodology as possible; this information will be very valuable for independent reviewers tracking patterns in the data, and it adds confidence to any inferences drawn from the analysis.

A policy of openness, transparency, and rigorous data collection is necessary if the results of parapsychology are to gain mainstream attention and respect; this can be accomplished by utilizing data registries and making available as much information about studies as possible.

Forced-choice ESP

To build our case for the forced-choice ESP paradigm, it is instructive to first review the central findings of the previous two meta-analyses on this approach, and then to identify consistent findings that suggest future research directions.

The Honorton-Ferrari Database

The first major meta-analysis of the forced-choice ESP approach was done by Honorton and Ferrari (HF; 1989). They meta-analyzed a heterogeneous database of 309 forced-choice precognition studies across 62 investigators between the years of 1935 and 1987. Studies were selected if significance levels and *ES*s could be calculated based on direct hitting. With nearly two million individual trials contributed by more than 50,000 individuals, they found an unweighted Stouffer's $Z = 11.41$, $p = 6.3 \times 10^{-25}$, and an unweighted mean $ES = 0.020$, with the lower 95% confidence estimate of the mean $ES = 0.011$. Moreover, 93 (30%) of the studies were statistically significant at the 5% level, which we calculate to have an exact binomial $p = 1.30 \times 10^{-14}$. The homogeneous database HF produced by removing outliers (using the "10% trim") yielded similar results: in 248 studies, the unweighted Stouffer's $Z = 6.02$, $p = 1.1 \times 10^{-9}$, and unweighted mean $ES = 0.012$, with the lower 95% confidence estimate of the mean $ES = 0.005$. In addition, 62 of the 248 studies (25%) were significant at the 5% level, which we calculate to have an exact binomial $p = 1 \times 10^{-14}$.To address file-drawer concerns, HF used Rosenthal's fail-safe *N* statistic and estimated that the ratio of unreported to reported studies to nullify the overall *Z* in the heterogeneous database would have to be 46:1. They also found a highly significant correlation between the study *z* scores and sample size, $r(307) = .16$, $p = .003$, two-tailed; by contrast one would expect, in the presence of publication bias, a negative correlation between *z* scores and sample size.

To address concerns about how study quality covaries with *ES*, HF coded[9] each study in terms of eight procedural criteria in the research reports (the criteria are reported in their publication). In plotting the quality ratings against *ES* for the heterogeneous database, they found no significant relation, $r(307) = .06$, $p = .29$, two-tailed. Likewise for the homogeneous database: $r(246) = .08$, $p = .20$, two-tailed).

[9] Though the coding was not blinded, HF note that the same not blinded coding method was used by Honorton in his 1985 ganzfeld meta-analysis, which yielded good agreement ($r(26) = .76$, $p = 10^{-6}$) with the independent "flaw" ratings of Ray Hyman.

Notable also is that HF found only weak evidence of experimenter effects in their heterogeneous database: the difference in the mean *ES* across investigators was barely significant, with $\chi^2(61) = 82.71$ and p =.03, two-tailed, and no evidence of experimenter effects in their homogeneous database, $\chi^2(56) = 59.34$, p =.36, two-tailed.

More striking are the moderating variables discovered in the homogeneous database. HF found that in the 25 studies using "selected subjects," individuals selected on the basis of prior performance in experiments or pilot tests produced a Stouffer's $Z = 6.89$ and mean $ES = 0.05$, with 60% of the studies significant at the 5% level. By comparison, the 223 studies using unselected volunteers had $Z = 4.04$ and mean $ES = 0.008$, with 21% of the studies significant at the 5% level. Moreover, the mean *ES* difference between selected and unselected participants was highly significant, t(246) = 3.16, $p = .001$.

Even more striking are the 17 "optimal studies" HF identified in the heterogeneous database. These are studies that used selected people along with trial-by-trial feedback. They produced a mean $ES = 0.12$ (by far the highest of any group of studies their database) and a combined Z = 15.84, with 15/17 (88.2%) studies significant at the 5% level. In the homogeneous database, there were eight optimal studies with a combined $Z = 6.14$ and mean $ES = 0.06$, with 7/8 (87.5%) studies significant at the 5% level. By comparison, there were nine suboptimal studies (which used unselected individuals and no trial feedback) in the homogeneous database, with a combined Z = 1.29 and mean $ES = 0.005$, with no studies significant at the 5% level. The mean *ES* difference between the optimal and suboptimal studies (in the homogeneous database) was also highly significant at $p = .01$. And contrary to skeptical expectations, the optimal studies had mean quality ratings significantly greater than the suboptimal studies[10]. These optimal studies constitute, in our view, the most interesting and potentially useful findings from the forced-choice ESP paradigm.

We now turn to the forced-choice ESP meta-analysis of Storm et al. (2012) to see how the HF findings have held up over time[11].

The Storm-Tressoldi-DiRisio Database

Storm, Tressoldi, and DiRisio (STDR, 2012) reported a meta-analysis of 91 studies of forced-choice design conducted by 96 investigators from 1987-2010. The studies were selected based on fulfillment of six study inclusion criteria, and masked quality ratings were assigned to each study using the same six criteria employed by HF[12]. An important difference from the HF meta-analysis is that only 36% of the studies tested precognition, the rest testing clairvoyance (48%) and telepathy (16%). The number of options, *k*, for possible targets in these studies also varied from $k = 2$ to $k = 26$.[13]

These 91 studies formed a heterogeneous database with a Stouffer's $Z = 10.82$, $p < 10^{-16}$ and an unweighted mean $ES = 0.04$. There were 8,123,626 trials and 221,034 hits. Using Stem-and-Leaf

[10] Optimal mean = 6.63, *SD* = .92; suboptimal mean = 3.44, *SD* =.53; $t(10) = 8.63$, $p = 3.3 \times 10^{-6}$

[11] We choose to exclude review of the Steinkamp's meta-analysis of forced-choice studies (1998) because most of the studies in her meta-analysis were already include in HF's, and those that weren't are already included in STDR's much more contemporary meta-analysis. A review of Steinkamp's findings can nevertheless be found in STDR.

[12] For the specific selection and quality criteria, see the section "Selection Criteria" for the former and "Procedure" for the latter in STDR.

and Box-and-Whisker plots to identify outliers, they formed a homogeneous database of 72 studies, forming a total of 790,465 trials and 214,513 hits. In this database, 35% tested precognition, 53% clairvoyance, and 12% telepathy. Telepathy studies produced the strongest effect (mean *ES* = 0.04), with the precognition and clairvoyance studies producing the same effect sizes (mean *ES* = 0.01). Overall, the homogeneous 72 study database produced a Stouffer's $Z = 4.86$, $p = 5.90 \text{ x } 10^{-7}$, and an unweighted mean *ES* = 0.01 with 95% *CI* = 0.01 to 0.02. Of these 72 studies, 16 (22%) reached significance at the 5% level, for which we calculate an exact binomial $p = 4.28 \text{ x } 10^{-7}$.

To address file-drawer concerns, STDR used as their primary analysis the Darlington-Hayes test, which is more conservative than the Rosenthal fail-safe statistic in that it allows all the "fail-safe *N*" studies to have negative *z* scores. They concluded that 187 unpublished studies with negative *z* scores would have to exist to nullify their 16 statistically significant studies. As a further check, they also used the Rosenthal statistic and found that 557 unpublished studies (or an unpublished to published ratio of nearly 8:1) would be needed to bring their Stouffer's *Z* to zero.

So far then, the STDR database is consistent with the HF database in finding statistically significant Stouffer's *Z*s and positive mean *ES*s for both their homogeneous and heterogeneous databases. They are also consistent in rejecting the file-drawer explanation. One may ask, however, if the fact that the STDR database contained mostly non-precognition forced-choice studies might confound these consistent findings. One may also ask how the overall results of the homogeneous STDR precognition studies compare to the results of the homogeneous HF precognition studies. As to the first, we note that STDR found no clear evidence of a significant performance difference between the telepathy, clairvoyance, and precognition studies in their database[14]. As to the second, we note that STDR found that their 25 precognition studies produced a mean *ES* = 0.01, Stouffer's *Z* = 1.92 and mean *Z* = 0.38; comparing them to the HF findings of a mean *ES* = 0.01, Stouffer's *Z* = 6.02, and mean *Z* = 0.38, they conclude that there is no substantial difference between the two sets of precognition studies.

A question left open by STDR however is how *z* scores correlate with study sample size. We found no correlation between these variables (*r* = .00). However, for all 35 precognition studies, we found a positive (though nonsignificant) correlation of *r*(33) = .21, *p* = . 24, two-tailed. Despite the nonsignificance of the correlation, which could plausibly be due to the small sample size, its magnitude replicates closely the correlation found by HF of *r*(307) = .16.

What about study quality versus *ES*? STDR found that quality ratings and *ES* values had a non-significant relationship of *r*(89) = .08, *p* = .45, two-tailed. This matches the HF findings of *r*(307) = .06, *p* = .29, two-tailed. On a related note, STDR found the relationship between quality ratings and study year to be positive and significant at *r*(89) = .25, *p* = .02, two-tailed, which is consistent with HF's finding of *r*(246) = .28, $p = 2 \text{ x } 10^{-7}$. STDR also found for their homogeneous database a highly significant incline of *ES* values by study year of *r*(70) = .31, *p* = .007, two-tailed. This last correlation did not match HF's, who had found a non-significant correlation between *ES* and study year of *r*(307) = .07, *p* = .21, two-tailed. However, STDR's subset of precognitive studies did yield a non-significant correlation between *ES* and study year: *r*(25) = .14, *p* = .26, one-tailed.

[13] Study design choice was not specified in the HF meta-analysis.

[14] They did find that telepathy and precognition approached a significant difference in mean *ES* (*p* = .09, two-tailed), but this of course can only be taken as suggestive of a real difference.

How about experimenter effects? Here STDR found that *ES* values were not significantly different between experimenter groups, $\chi^2(15, N = 57) = 18.10$, $p = .36$, two-tailed. This is consistent with HF's findings for their homogeneous database, $\chi^{2}(56) = 59.34$, $p =.36$, two-tailed.

What about moderator variables? STDR did not investigate moderator variables in their database. However, we found 11 studies that used selected individuals, producing a mean *ES* = 0.09, with 8/11 (73%) significant at the 5% level. By comparison, the 80 studies using unselected people produced only 21/80 (26.3%) significant studies, with mean *ES* = 0.03. In contrast to HF's findings, we found that the mean quality rating of the studies using selected participants was lower than the mean quality rating for studies using unselected ones (mean $q = 0.69$ vs. $q = 0.81$, respectively), but the difference between ratings was not significant[15]. In addition, 6/11 of the selected participants studies produced a mean quality rating of 0.86, with no study rated less than 0.80, and yet the mean *ES* = 0.08 for these six studies is more than double the mean *ES* of the unselected participants studies. Overall, then, these findings of a significant performance difference between studies using selected and unselected individuals in the STDR database are consistent with those of HF for their comparison of studies using selected vs. unselected volunteers.

Given the significant *ES* difference between selected and unselected subjects studies, we hypothesized that the 11 selected studies would produce a significantly stronger correlation between *z* score and sample size than the unselected studies. This hypothesis was confirmed, as we found that these 11 studies produced a strong correlation of $r(9) = .64$, $p = .01$, one-tailed. By comparison, the 80 unselected studies produced a null correlation ($r = 0$). Although HF did not do this analysis for their selected and unselected studies, these robust findings suggest it would be worthwhile to do it for the HF database as well[16]. Unfortunately, there was insufficient information in the STDR database to identify all the selected participants studies that used trial-by-trial feedback or no trial feedback, so neither optimal nor suboptimal studies in the STDR database could be identified. Moreover, there were only three precognition studies that used selected individuals, preventing us from making direct and reliable comparisons with the selected precognition studies in the HF database.

Taking into account the consistency of findings between the HF and STDRs databases, and the analyses that have yet to be done, we now turn to making recommendations for future research directions.

Future Directions

In our view, the most impressive findings of the HF and STDR meta-analyses are the consistent results of the selected individuals studies in both databases, and the results of the optimal studies in the HF meta-analysis. There is clearly a very significant performance advantage—both in the *ES* difference and the proportion of independently significant studies—for these studies in contrast to studies using unselected individuals and suboptimal conditions.

[15] Selected subjects studies mean rating = .69, *SD* = .23; Unselected subjects studies mean rating = .80, *SD* = .21; $t(89) = 1.61$, $p = .11$ (two-tailed).

[16] Derakhshani has attempted to acquire the data from the studies in the HF database in order to do this analysis. Unfortunately, Honorton is deceased, Ferrari seems to have lost contact with the field, and no other researchers in parapsychology seem to have access to the HF database.

In terms of finding a parapsychological research paradigm that offers the best chance of feasibly and consistently producing high replicability rates (and without the complication of experimenter effects), the forced-choice approach seems to be the most promising. Let us illustrate this with a simple power analysis. Consider that in the STDR database, the mean sample size of the 11 studies with selected individuals is 3,077 trials. Conservatively using the *ES* = 0.055 value for the eight optimal studies in the HF database, the statistical power of a 3,077 trial optimal study would then be 92% (using an alpha level of .05). Considering our finding of a strongly positive and highly significant correlation between *z* scores and sample size for selected studies in the STDR database—a trend that would be theoretically expected under power analysis assumptions—and the similarly high proportion of independently significant selected studies in both the STDR and HF databases, we can reasonably expect our power analysis result to be a reliable indicator of the probability of a significant study outcome under the assumed conditions. Thus, for would-be replicators of the forced-choice ESP effect, we make the following recommendations:

1. Whenever possible, make exclusive use of selected individuals (e.g. individuals selected on the basis of prior performance in experiments or pilot tests).
2. Whenever possible, implement trial-by-trial feedback to take advantage of the improved replication rate reported with that method.
3. Preplan the number of trials in a study on the basis of a power analysis to achieve at least 90% power (it is recommended to use a conservative *ES* estimate such as the one used in our example above).
4. Use a precognition design with true random number generators to select the targets, so as to reduce the possibilities of sensory leakage and anticipation bias.
5. Preregister the study in any one of the registries available to parapsychology researchers (Open Science Framework, 2014; KPU Registry, 2014).
6. Log as much information about participants and methodology as possible; this information will be very valuable for independent reviewers tracking patterns in the data, and it adds confidence to any inferences drawn from the analysis.

With these recommendations implemented, we anticipate (assuming the validity of the empirical findings on which these recommendations are based) a vast improvement in the overall quality, *ES*, and replicability rates of future forced-choice ESP studies. We also expect that if the proportion of independently significant studies in a future forced-choice meta-analytic database could approach 90%, as our analyses suggest, parapsychology would have yet another research paradigm that would compel skeptical mainstream scientists to attempt replications. Conversely, if after implementing these recommendations in future replications, results fail to improve as predicted or even to replicate the HF and STDR findings, there would be serious doubt about the validity of the empirical findings involving selected individuals and optimal studies. It would then be appropriate to ask and try to understand why the previous forced-choice meta-analyses produced the seemingly impressive results they did.

Remote Viewing and Non-ASC Free-Response Studies

In this section, we review the findings of the remote viewing paradigm and the non-ASC free-response paradigm. We also compare the two paradigms and show that they are significantly different from each other in non-trivial ways (despite RV often being classified as just another form of non-ASC free-response). Suggestions for fruitful future research directions are also given.

Remote Viewing

Studies involving remote viewing (RV) have had, and perhaps still have, major popular appeal, partly because some of the experts who participated in the early remote viewing projects at the Stanford Research Institute (SRI), starting in the seventies, and at the later military-sponsored Science Applications International Corporation (SAIC), have received considerable media attention (e.g., Joe McMoneagle). In the classic version of RV, the procedure requires an agent to transmit what he or she is seeing in some location to a masked recipient in some other distant location (typically several kilometers away, but sometimes much more). The recipient is instructed to use special visualization techniques in a normal (apparently non-altered) state of consciousness, and usually also to work with a blinded experimenter interviewing him or her in real-time. Many RV sessions are also done with a precognition design, in which the target location is chosen after the remote viewing session is completed. In each design, the agent usually knows the recipient. A more extensive narrative review of the history of RV and projects related to RV—for instance the Mobius Project, a famous use of remote viewing for archeology—is given by Schwartz (2014).

For examination of the early work, a comprehensive evaluation of the SRI and SAIC research was provided by statistician Jessica Utts (1996), who was part of an independent commission called upon to express a judgment on the RV programs, along with mainstream counter-advocate Ray Hyman and other experts. Table 119 shows the results obtained in the two research centers. A curious and noteworthy finding of Utts' analysis is her comparison of the performance of experienced and novice SRI viewers to experienced and novice ganzfeld receivers at PRL. Utts found that the ESs of experienced SRI viewers and experienced PRL receivers were nearly identical (0.385 vs. 0.35), as were the ESs of novice SRI viewers and novice PRL receivers (0.164 vs. 0.17). This would seem to suggest that RV participants with comparable characteristics to ganzfeld participants perform comparably well. It would also seem to suggest that the putative psi-mechanism responsible for RV is also responsible for ESP in the ganzfeld condition. From the viewpoint of the widely-accepted noise-reduction model of psi ability, however, this seems puzzling —if the noise-reduction model is correct, one would expect the PRL receivers (who were necessarily placed into an altered state of consciousness via the ganzfeld procedure) to produce significantly better ESs than the SRI remote viewers (who supposedly used normal states of consciousness). One of us (Derakhshani) has examined this issue in detail and found evidence to suggest that, contrary to the standard description of RV as a non-altered state free response protocol, the experienced viewers at SRI and SAIC did in fact employ an altered-states condition. In particular, the distinguished remote viewer, Joe McMoneagle (personal communication, 2014), has indicated that he himself did use a self-induced altered-state condition during his RV trials with SRI and SAIC. He describes it as "a complete and full disassociation of what is going on around me, while I'm attempting to collect material on an assigned target." He also noted that "Almost without exception, in the first half of our program, the excellent remote viewers were all learning to control their own ability to disassociate from whatever is going on in their life at the time of their remote viewings." Thus it would seem that, at least the "excellent" viewers at SRI (virtually) universally employed an altered states condition. While this helps to reconcile Utts' findings with the noise-reduction model, it does not entirely do so. It is unclear to what extent the novice viewers at SRI employed the "disassociation" altered state (if at all), and likewise for the experienced and novice viewers at SAIC. In fact McMoneagle also commented that, "Novice remote viewers do not necessarily understand how to disassociate from what is going on around them while thinking about a remote viewing target, or that they must in order to reduce the noise to details process." If novice viewers did use altered states conditions, then there would seem to be no inconsistency with the noise-reduction model. But if they did not, it would indicate that the noise reduction-model is inconsistent with the SRI RV results, and therefore either incorrect or incomplete. So this remains an open question for further study.

Another important source of evidence for RV studies comes from the PEAR laboratory (Princeton Engineering Anomalies Research), run by Robert Jahn and Brenda Dunne—now closed, but succeeded by the International Consciousness Research Laboratories (www.icrl.org). Jahn and Dunne (2003) provide a summary of 25 years of research from the PEAR project, consisting of 653 tests conducted with 72 volunteers. We report their results in Table 8.

The protocols used at the PEAR lab had important discrepancies: in some experiments, individuals were completely free to express the general sense and details of their impressions, while in others they had to make use of structured questionnaires within which they indicated the presence or absence of pre-listed attributes. Their work had, therefore, characteristics of both free-response and forced-choice ESP protocols. Notably, the free-response scenario was more productive, achieving a 67% success rate, while the forced-choice protocol provided a success rate of only 50%—at chance. Interviews of the recipients who had used questionnaires showed that they may have felt limited in their perceptions, switching from introspective modes of thought to analytical ones. According to Jahn & Dunne, the successful transmission of psi information requires the presence of a certain amount of uncertainty and/or a nonanalytic mental state. Under conditions of waking consciousness, better results were obtained when recipients were allowed to say everything that came into their minds rather than when they were forced to constrain their experience into a narrow range of options (One might also wonder if percipients in PEAR's studies employed altered-state conditions like disassociation. To the best of our knowledge, this was never assessed. So PEAR's RV studies at present cannot be used to test the noise-reduction model). Jahn and Dunne also reported that there was no difference in the success rate of trials for which the target was located close, as opposed to far (from a few meters to a few kilometers).

Table 8: Summary of evidence related to all available studies related to RV.

Source	Trials	Hits	*ES* (*z*√*N*)	95% *CI*
SRI database (Utts, 1996)	770	262	0.20	0.17-0.23
SAIC database (Utts, 1996)	445	160	0.23	0.19-0.27
Milton (1997)	2682	798*	0.16	0.10-0.22
Dunne & Jahn (2003)	653	223*	0.21	0.18-0.24
Bierman & Rabeyron (2013)	550	349	0.27	0.23-0.31
1994-2014	314	118	0.39	0.14-0.64

SRI =Stanford Research Institute; SAIC = Science Applications International Corporation;; *estimated from Stouffer's *Z*;

All these findings notwithstanding, the PEAR RV studies do need to be qualified in important ways. For the 653 tests conducted, there were two distinct protocols used for target selection. For 211 trials, the "volitional mode" was used (outbound agents could freely choose any targets they liked), while 125 were in the instructed mode (targets were randomly selected by a well-calibrated REG). Hansen, Utts, and Marwick (1991) argued that the overall effect size and significance level calculated for the volitional trials are meaningless because, by not having targets randomly selected, there is no theoretical or empirical baseline distribution to which the observed data can be compared. Dobyns, Dunne, Jahn, and Nelson (1992) retorted that, nevertheless, the instructed trials produced an overall large effect size of 0.516 with overall z = 5.77. However, another objection Hansen et al. (1991) raise is that both the volitional and instructed trials used sampling of targets without replacement. Sampling without replacement is well known to lead to violations of statistical independence between trials, which the PEAR analyses did not take into account, and which can lead to p values incorrect by several orders of magnitude. Dobyns et al. responded to this critique with detailed statistical analyses to demonstrate that, on the whole, sampling without replacement could not be used to invalidate the overall effect size and statistical significance of their database. In our view, the difficulty with PEAR's response is that unless someone takes the time to work through their analyses and is personally convinced of their conclusions, it is difficult to trust that the results for the instructed trials are as reliable and valid as, say, the (later) SRI and SAIC studies, where sampling without replacement was a non-issue. In any case, it seems clear that PEAR could have done better with their studies simply by always using sampling with replacement and always randomly selecting targets. Doing so would have made their RV studies impervious to these two potentially fatal criticisms.

Another recent source of evidence related to a specific RV procedure—Associative Remote Viewing (ARV)—has been evaluated by Bierman and Rabeyron (2013). With this procedure, the remote viewer describes a target; a picture to be chosen in the future. The silver market or another indicator is then used as a random number generator to select the target. If the viewer correctly describes the target, by implication (or association) the viewer "describes" the silver market, although the viewer might be totally unaware of this setup. The summary of their comprehensive review of 18 studies is presented in Table 8.

To complete the available evidence, we retrieved all (nine) studies completed up to February 2014, and not included in the previous databases. Some of them were included in the meta-analysis of Storm et al. (2010), under the category of non-ASC free-response studies in a normal state of consciousness. The summary is presented in Table 8.

Looking at the *ES*s and their confidence intervals presented in Table 8, the results appear to be quite homogeneous. Considering that the older databases are made of experiments carried out in the 1970s, and the more recent databases are of experiments carried out in the 2000s, we can affirm that there is no sign of decline in almost 40 years. This trend is best visualized in the Figure 1.

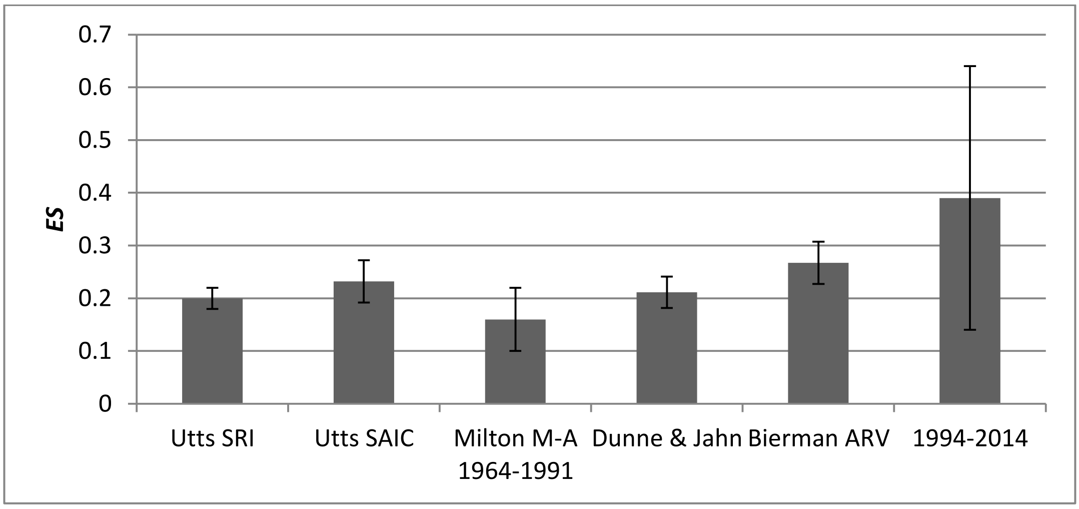


Figure 1. *ES* with corresponding 95% confidence intervals related to the different databases in chronological order.

The larger confidence intervals in the 1994–2014 database are due to the small number of studies and a large variability of the *ES*s, ranging from -0.03 in Subbotsky and Ryan’s (2009) Study 2 to 0.93 in Targ (1994).

To the best of our knowledge, the analyses presented here constitute the closest thing available to a comprehensive meta-analysis of RV studies throughout the past 40 years. We find this is regrettable. Such a meta-analysis could potentially tell us what the moderator variables are for RV (e.g. selected vs. unselected participants, target characteristics, etc.), how RV *ESs* correlate with study sample sizes, how RV study quality correlates with *ES*, and how RV study characteristics compare to ASC and non-ASC study characteristics. This information would be useful to know for would-be replicators of RV studies, as well as for improving our understanding of how RV facilitates psi ability as compared to other techniques.

In any case, based on our analyses, we can make the following observations that may be useful for would-be replicators of RV. For the *ES* of 0.23—the unweighted average *ES* between SRI, SAIC, and Bierman & Rabeyron—the sample size needed to reach 80% power at the 5% level is 55[17]. Since SRI found that experienced viewers performed significantly better than novice viewers, and since SAIC used many of the same viewers as SRI, we can be reasonably confident that this performance difference carried through to SAIC. And if we expect that RV is a consistent phenomenon, then we can be reasonably confident that this performance difference is a general characteristic of RV. Accordingly, for experienced viewers (*ES* = 0.385), the sample size needed to reach 80% power is only 33, while for novice viewers (*ES* = 0.165) it is 77. Clearly experienced viewers are more advantageous when attempting to replicate RV, and this is prior to considering any other factors (e.g. target characteristics) which could further amplify *ES* and thereby further lower the number of trials needed to have a high probability of a successful study.

Non-ASC Free-Response Studies

As mentioned already, RV is often assumed to be a non-ASC free-response protocol (even though we have seen that this is far from clear). As a result, RV studies are often mixed with other non-ASC free-response studies in meta-analyses. These other types of non-ASC studies will typically use receivers in a normal, passive, waking state of consciousness (e.g. sitting quietly with eyes closed) rather than using RV techniques to actively identify a target. The first such meta-analysis was carried out by Julie Milton (1997). She meta-analyzed 78 studies covering the period from 1964 to 1992, conducted by 35 different authors and with a total of 2,682 tests conducted with 1,158 individuals. Her goal was to test the noise-reduction model by comparing the overall effect produced by her database with that of ganzfeld meta-analyses. The summary of Milton's meta-analysis is reported in Table 8. While Milton clearly obtained evidence for an effect, and while the magnitude of this effect was actually larger than the overall effect obtained in the latest ganzfeld meta-analysis at the time (the Milton-Wiseman meta-analysis), it is unclear to what extent the overall effect in her database depended on the RV studies specifically. In addition, it has been pointed out by Storm et al. (2008) that Milton actually included 25 studies that used meditation, mental imagery training, relaxation, and ganzfeld, among other ASC studies. This is perplexing given the objective of her meta-analysis and seems to belie her conclusion that the results contradict the noise-reduction model.

Thirteen years later, Storm et al. (2010) performed an independent and more up-to-date meta-analysis of non-ASC vs. ASC free-response studies, in order to test the noise-reduction model. They compared three databases – ganzfeld, non-ganzfeld ASC, and non-ASC free-response. Their non-ASC database was initially composed of a heterogeneous set of 21 studies (unweighted mean *ES* = 0.102, Stouffer *Z* = 2.17). Upon removing seven outliers (identified with Box and Whisker plots), they formed a homogeneous dataset of 14 studies that showed no evidence for a real effect (mean *ES*s = -0.029, Stouffer *Z* = -2.29). In addition, an exact binomial test on trial counts (*N* = 1,872, hits = 455) yielded only a 24.3% hit rate (binomial *z* = -0.67), where 25% was the null hit rate. By contrast, they found clear evidence for real effects in their homogeneous ganzfeld database of 29 studies (mean *ES* = 0.142, Stouffer *Z* = 5.48) and their homogeneous non-ganzfeld ASC database of 16 studies (mean *ES* = 0.110, Stouffer *Z* = 3.35). Storm et al.'s findings for their non-

[17] To the best of our knowledge, no regression analysis has ever been undertaken to see how $\sqrt{N}$ correlates with *z* for RV studies. Nevertheless, if it is the case that the underlying psi mechanism is the same for RV as it is for ganzfeld (as Utt's analysis would seem to suggest), it seems plausible to expect that similar correlations will be found for RV studies using similar types of participants. Consequently, we should have a similar level of confidence that RV studies under the prescribed conditions will yield replication rates comparable to these power percentages.

ASC database must also be qualified, however. The fact that their non-ASC database was extremely heterogeneous suggests that there was a real effect produced in at least some of the studies. Arguably, the better approach would have been to apply a random effects model to the 21-study database in order to estimate the overall ES with the heterogeneity taken into account. In addition, the heterogeneity suggests that there may be moderator variables for the overall effect. Looking for these moderator variables and blocking the studies accordingly could then suggest an explanation as to what non-ASC study characteristics facilitate psi better than others. So for example, perhaps RV studies produced significant and similar effect sizes but the other non-ASC studies did not. In fact, Storm et al.'s database contained three RV studies that produced similarly large *ESs* (0.582, 0.646, and 0.723 respectively) which were independently significant (*z*'s = 1.84, 4.57, and 3.54 respectively) and which were among the seven outliers removed. Removing only these three studies from the database brings the mean *ES* down to a non-significant level (N-weighted mean *ES* = -0.0037, Stouffer *Z* = 0.00), and none of the remaining studies are independently significant. In addition, Cochran's *Q* (the chi-square test used also in the ganzfeld section) on the remaining 18 non-ASC studies shows no indication of the presence of heterogeneity (*Q* = 11.63, *p* = 0.18), and likewise $I^2 = 0$. So it would seem that something about the RV condition is a strong moderator variable for this database. That is, something about the RV condition is considerably different from other non-ASC conditions in terms of facilitating psi ability. Understanding what, precisely, are the factors in these three RV studies that produced such large *ESs* (e.g. perhaps viewers using self-induced disassociation states, targets with large Shannon entropy gradients; May, 2011; consensus judging methods; Storm, 2003; large experimenter effects, etc.) in contrast to the other non-ASC studies remains an open question for further research.

What seems clear is that RV studies are their own category of free-response studies that don't easily classify into either non-ASC or ASC, and yet produce *ESs* at least as good or better than the most successful ASC studies. What also seems clear is that non-ASC/RV studies don't seem to facilitate psi at all in a laboratory setting, and can thereby serve as a useful control/comparison for ASC and RV studies.

Future Directions

Even with all the variations in RV protocol, the accumulated evidence from SRI to the present day seems to consistently demonstrate the existence of a real RV effect. Nevertheless, we will summarize our general recommendations for how to consistently obtain larger effects and more reliable results:

1. Participants should be trained (i.e. experienced) in the RV technique before running a study.
2. Participants should be allowed to freely report their perceptions, rather than being limited by analytical questionnaires or guidelines.
3. Participants should be requested to complete only one or two trials a day, to prevent fatigue and boredom.
4. Participants should receive feedback both during and after each trial.
5. The choice of the characteristics of the target and the decoys is important for enhancing *ES*. For example, May (2011) studied and released a pool of images that were classified for their information entropy, that is the amount of information they represent. He found that target

images with large Shannon entropy gradients relative to the decoys correlated strongly with *ES*. Therefore, targets and decoys used in future RV studies should emulate this finding.

6. The choice of protocol for data collection and analysis is also relevant to enhancing *ES*. May, Marwaha, and Chaganti (2011) note that the use of rank-order assessment (to identify the target) and the "fuzzy set technique" (to assess the quality and reliability of the viewer's mentation to the target) have produced a substantial number of significant results in the laboratory as well as in various application environments. Therefore, future RV studies should continue to make use of these two protocols.
7. Log as much information about participants and methodology as possible; this information will be very valuable for independent reviewers tracking patterns in the data, and it adds confidence to any inferences drawn from their analyses.
8. Plan for the power of future studies to be at least 80%. Using experienced participants (predicted *ES* = 0.385) comparable to those used at SRI, this would correspond to at least 33 trials. Using novice participants (predicted *ES* = 0.165) comparable to those at SRI, at least 77 trials.
9. Implement a strict policy of using trial registries for all trials conducted.
10. Log as much information about participants and methodology as possible; this information will be very valuable for independent reviewers tracking patterns in the data, and it adds confidence to any inferences drawn from the analysis.

There are also a number of open research questions about the moderator variables for RV, the replication characteristics of RV studies (e.g. how $\sqrt{N}$ correlates with z), the specific characteristics of the RV technique that (apparently) facilitate psi in contrast to non-ASC free-response studies, etc. The answers to these questions await a comprehensive RV meta-analysis.

Dream ESP

We first review the results of the Maimonides studies and then the meta-analytic findings of Sherwood and Roe (2003) and STDR (2010), as well as the recent work of Watt from the KPU (2014). Our objective is to identify study characteristics that can boost effect sizes and replicability rates in future dream studies, and to argue that this is a research paradigm worth revisiting.

The Maimonides Studies

The Maimonides Dream ESP (DESP) laboratory, which ran from 1962 to 1978 at the Maimonides Medical Center in Brooklyn, was the first major attempt to study the possiblity of DESP under controlled laboratory conditions (Ullman, Vaughan, & Krippner, 2003). Throughout its duration, 13 formal studies were conducted along with three groups of pilot studies. Of the 13 formal studies, 11 were designed to study telepathy, and 2 precognition; for the three pilot sessions, clairvoyance, telepathy, and precognition were studied. Since many thorough reviews of the Maimonides methodology already exist (Child, 1985; Sherwood and Roe, 2003; Ullman et al., 2003; Krippner and Friedman, 2010), we will refer the interested reader to them for details. What we will note, however, is that despite several criticisms raised about the Maimonides methodology, most have been shown to be unfounded; and those that are valid have been shown to not compromise the obtained overall results in any significant way (Child, 1985; Sherwood & Roe, 2003; Krippner & Friedman, 2010).

The first meta-analytic review was conducted by Irvin Child (1985), who observed that the only way the Maimonides study results could be meta-analyzed was in terms of binary hits and misses. This yielded 450 DESP trials (based only on the results of independent judges), with an overall hit rate of 63% (where mean chance expectation was 50%), and exact binomial probability of 1.46×10^{-8} or odds against chance of around 75 million to 1. Moreover, 20 of the 25 sets of data analyzed were above mean chance expectation. These impressive overall results obtained under the highly stringent Maimonides protocols have motivated numerous replication attempts from 1977 to the present day[18].

The Post-Maimonides Studies

Here we review the findings of three major post-Maimonides studies—the meta-analytic review of Sherwood and Roe (2003), the meta-analysis of STDR (2010), and the Perrott-Warrick Dream ESP study recently completed at the KPU by Caroline Watt (2014).

Sherwood and Roe

The first meta-analytic review of the DESP studies after Maimonides was conducted by Sherwood and Roe (SR; 2003). At that time, they had identified 23 formal reports of DESP studies published prior to 2003. In contrast to the Maimonides studies, which mainly studied telepathy, only nine of the post-Maimonides studies did so, with 13 evaluating clairvoyance, and 4 precognition. (Some used a combination of telepathy and clairvoyance or precognition and clairvoyance.) The presumed rationale for this greater emphasis on the clairvoyance approach was greater methodological simplicity, in that it does not require the use of a sender. Also, whereas the Maimonides studies used independent masked judges, the post-Maimonides studies used participant and experimenter/sender judging. However, they followed the Maimonides studies in using consensus rank-judging procedures to calculate effect sizes when possible.

Of these 23 studies, 21 reported sufficient information to obtain an outcome measure. However, because many of these studies used different outcome measures (or sometimes more than one), their results could not be immediately compared or meta-analyzed. To do this, SR applied a common effect size measure, namely, $ES = z/\sqrt{N}$. The range of *ES*s was from -0.49 to 0.80, which strongly suggests a heterogeneous *ES* distribution (although no test of heterogeneity was conducted). They also found that the most successful post-Maimonides studies were conducted by particular research groups. For example, the most successful studies were two telepathy experiments conducted by Child (1977) (*ES* = .58 and .80 respectively). Other research groups (Dalton et al., 1999; Sherwood et al., 2000; Dalton et al., 2000) produced a series of successful clairvoyance studies (*ES* from .24 to .63). From these results SR concluded that replications have been possible across laboratories and groups of researchers. To assess whether independent

[18] It should be added that while the Maimonides lab was in operation, eight conceptual replications were attempted, five of them by independent researchers. In an extensive analysis of these eight studies, Sherwood and Roe (2003) note that two demonstrated successful effect size replications of individual performances within the Maimonides program; three of the replication attempts were difficult to evaluate due to limited amount of details available in published reports. Nevertheless, of the remaining six replication attempts, they concluded that none could be considered successful. They do note, however, that these unsuccessful conceptual replications deviated in protocol from the Maimonides studies in important ways that might well have influenced the outcomes.

replication across laboratories and researchers was also achieved, they compared the range of results of telepathy, clairvoyance, and precognition studies. The most successful studies were clairvoyance (*ES* = -.49 to .63, median *ES* = .25) and telepathy (*ES* = -.27 to .80, median *ES* = .10), with precognition being least successful (*ES* = -.34 to .07, median *ES* = .04).

To compare the post-Maimonides overall results to those from Maimonides, the latter were converted to the same *ES* measure. Calculating a combined effect using Fisher's transformed values of *r*, the Maimonides studies produced an overall *ES* = 0.33, 95% *CI*s [0.24, 0.43]. The post-Maimonides studies produced an overall *ES* = 0.14, 95% *CI*s [0.06, 0.22]. The Maimonides overall *ES* was significantly greater than the post-Maimonides one, *t*(34) = 2.14, *p* = .04, although only marginally so. Although no coding or comparison was done of the quality of the Maimonides studies and the post-Maimonides studies, several differences in methodology were noted that could explain the difference in overall *ES*s.

For example, SR noted that only one post-Maimonides study used laboratory monitoring of EEG or deliberate awakening from REM sleep for recording dream recall, although awakening of participants during REM sleep is advantageous in that dream recall is much more likely and can lead to more detail and longer overall reports. Indeed, reviews of studies using laboratory awakening from REM show that dreams are reported in about 75%–80% of cases. By comparison, spontaneous awakenings in the morning are less likely to lead to dream recall and any dreams that are reported tend to be from the last REM period only. This suggests that judges in the Maimonides procedure probably received more detailed dream recollections upon which to base their judgments.

Another potential advantage of the Maimonides methodology is that, in the telepathy studies, sending efforts were synchronized with REM periods, whereas in post-Maimonides research the sending period was less systematic. In addition, the majority of post-Maimonides studies involved participants sleeping in their own homes instead of a laboratory. They also differed in that the Maimonides studies used independent masked judges whereas post-Maimonides studies used participant judging. It is possible that some experienced judges may have been better able to discern between normal dream information and ESP-influenced dream information. SR also observed that some of the post-Maimonides research suggests that consensus judgments (as were used in Maimonides) may offer a slight advantage over individual judgments (as were used in most post-Maimonides research).

In addition, SR observed that the Maimonides studies went to great lengths to screen for effective senders and receivers (including recruitment of participants with prior success in ESP studies) and using psi-conducive pairings. By contrast, post-Maimonides studies did not screen so carefully or use psi-conducive pairings.

Another potentially important difference was that the majority of the Maimonides studies investigated telepathy, while the majority of post-Maimonides studies investigated clairvoyance or precognition. If the sender plays an active role in the ESP process, or even if having a sender simply induces a more comfortable and optimistic mindset in receivers, this would of course have an impact on the likelihood of outcome success.

Yet another important difference SR noted is that the Maimonides team chose visual targets because of their emotional intensity as well as their vividness, color, and simplicity—factors that were regarded as a crucial feature of the Maimonides protocol. By contrast, the post-Maimonides research did not typically select target pools using these criteria.

Finally, SR noted that a number of studies have found that the earth's geomagnetic field (GMF) and local sidereal time (LST), seem to correlate significantly with the success or failure of DESP trials. In particular, periods of lower GMF activity have been associated with reports of spontaneous precognitive dreams (Krippner et al., 2000) and greater accuracy in experimental dream ESP trials (Dalton et al., 1999; Krippner and Persinger, 1996; Persinger and Krippner, 1989, Sherwood et al., 2000). The work of Spottiswoode (1997) also showed that LST might mediate the relationship between GMF and free-response ESP performance more generally. This finding is also supported by the more recent work of Ryan (2009). These factors were also not taken into account in the Maimonides studies or in most of the post-Maimonides studies.

Storm-Tressoldi-DiRisio

In addition to the SR meta-analytic review, an updated database of post-Maimonides studies can be found in Storm et al. (2010). These involved eight post-Maimonides studies between the years 1999 and 2007. With the exception of the three studies in 2007, the rest were already part of the SR database. Storm et al. (2010) did not meta-analyze these studies directly but rather combined them with other non-ganzfeld altered-states studies (e.g. hypnosis and meditation studies). Unlike the SR database, STDR also included study quality ratings for all eight studies. These eight studies combined produced an unweighted Stouffer's $Z = 2.38$ and unweighted $ES = 0.149$. The range of mean study quality ratings was from 0.75-0.85, out of a possible 1.00. Unfortunately, it would not be useful to plot these quality ratings versus *ES* given that 7/8 studies had the same quality rating of 0.85, with the other having 0.75. Nevertheless, other analyses can be considered.

We combined the three 2007 studies with the SR database and used the more straightforward method of computing an unweighted overall *ES* for these studies. The original SR database was found to yield an overall $ES = 0.19$. With the three 2007 studies included, the overall *ES* drops slightly to 0.17, still well above chance expectation. Unfortunately, we are unable to plot *z*-score vs. sample size in the combined database because several of the studies in the SR database reported t-values but not *z*-scores. Nevertheless, we were able to conduct a regression analysis for the STDR database alone. For these eight studies, the correlation is linear with $r = .46$ ($p = .13$, one-tailed). Although not statistically significant, this correlation is strong and suggestive that the combined database would yield a significant correlation. A regression analysis that we are actually able to perform for the combined database is *ES* vs. study sample sizes (which range from 5 to 100 trials). This produces a weakly negative linear correlation of $r = -.16$, which is not significant, $t(21) = 0.75$, $p = .46$, two-tailed)[19]. Although this correlation is negative, it is small in magnitude and the fact that it does not reach statistical significance suggests it is consistent with chance fluctuations (and, incidentally, consistent with little to no file drawer). This is also consistent with what statistical power assumptions would predict for a real effect. In sum, this finding for the combined database, along with our plot of *z*-scores vs. sample size for the eight STDR studies, is suggestive that the post-Maimonides studies conform to statistical power assumptions.

[19] One study (Hearne, 1981b) had to be excluded because no sample size was reported. This study reported an *ES* of exactly zero.

Watt’s Precognitive Dream Study

The most recent DESP study efforts have come from the KPU, led by Caroline Watt (2014). From October 2010 to December 2014, The Perrott-Warrick Fund appointed Watt as their Senior Researcher for a program of research into the psychology and parapsychology of precognitive dream experiences. Watt’s study recruited individuals online through Twitter to participate in an online precognitive dream ESP study. Upon filling out questionnaire’s about their beliefs in precognitive dreaming and precognitive dream experience, it was found that a majority of the recruited participants held a belief in precognitive dreaming (66%) and that a majority reported having a precognitive dream they considered evidential (72%). (In other words, a majority of the recruited participants could be characterized as selected, according to Storm et al.’s (2010) criterion.) Participants were asked to take note of their dreams over 5 mornings (which they were asked to remind themselves, before sleeping each night, would be linked to a target clip that they would later view), after which they were sent a questionnaire asking for an anonymous summary of their week’s dreams. Subsequently, they were sent a link to watch the ‘target’ clip (posted on YouTube). (Note that participants never saw the decoy clips.) Participants were also asked to fill out a 100-point similarity rating questionnaire to rate how similar they felt their dream content, themes, and emotional tone were to the target clip.

Two independent judges, who did not know the identities of the participants (and vice versa), were then asked to apply similarity ratings between the dream summary contents and the video clips, and use these ratings to rank-order (from greatest to least similarity) the four clips (one of which was the target) for each of the dream summaries of the participants. A rank-1 match of the target clip to a dream summary would then be a direct hit. The precognitive aspect of the study design refers to the fact that the target clip for any one trial was randomly selected (by a true random number generator from RANDOM.ORG) and sent to participants only *after* the independent blind judges had submitted their rank-orderings for that trial.

In 200 trials (pre-planned as four trials each from 50 participants), 64 direct hits were obtained for a 32% hit rate and *ES* = .16. Using an exact binomial test, this hit rate is highly significant (z = 2.21, p = 0.015, one-tailed) and the *ES* is nearly identical to the unweighted mean of the other post-Maimonides DESP studies (*ES* = .17). A notable difference, however, is that the post-Maimonides precognitive DESP studies in SR’s database produced the smallest *ES’s* (*ES* = -.34 to .07), in comparison to telepathy and clairvoyance designs.

It is also worth mentioning that, apart from informal pilot trials conducted as practice for the formal trials, Watt has stated (personal communication) that there are no unreported (formal) trials. So Watt’s results constitute a file-drawer free dataset.

Curiously, Watt (personal communication) has expressed that she does not believe that the results of her study support the precognitive psi hypothesis, because, while the obtained hit rate is statistically significant, participants’ similarity ratings of their dream summaries to the target clips did not significantly differ whether they were rank-1 matches or rank-2-4 matches (as determined by the independent judges). Watt believes that if the psi hypothesis is correct, then this should not have been the case.

However, in our view, this finding is not necessarily inconsistent with the psi hypothesis (e.g. defined by Bem and Honorton (1994) as "anomalous processes of information or energy transfer, processes such as telepathy or other forms of ESP that are currently unexplained in terms of known biological mechanisms."). It could well be the case that participants needed to make similarity ratings between their dream summaries and all four possible target clips, in order to notice relevant differences between their summaries and the clips and identify the clip with the most similarities. Alternatively, if one is concerned about the possibility that participants might precognize one of the decoy clips, one could instead leave it to the independent judges to apply the similarity ratings to all four possible target clips. As noted above, this is exactly what the judges did, and they were able to identify the target clips at a rate (32%) significantly greater than chance expectation (which is clearly consistent with the psi hypothesis).

In sum, Watt's precognitive DESP study is especially interesting in that it is the single largest post-Maimonides study ever conducted, it produced a statistically highly significant hit rate, an *ES* nearly identical to the mean *ES* of all the other post-Maimonides DESP studies (and considerably larger than the *ES* range of the other precognitive DESP studies), was exceptionally methodologically rigorous by using a precognitive design and being an online study, had no unreported formal trials, and the lead experimenter (Watt) is avowedly skeptical of the results of her own study. Taken together, these observations make the post-Maimonides DESP paradigm look particularly promising for follow-up research. A logical starting point would be for researchers to independently (exactly) replicate Watt's study.

Future Directions

Although the existing analyses of the Maimonides and post-Maimonides studies clearly indicate a non-chance effect that excludes mean chance expectation, a great deal remains to be understood about the DESP effect. Here we list open research questions about the full post-Maimonides studies:

1. How do study *z*-scores correlate with study sample sizes?
2. How does study quality (e.g., using the Storm et al., 2010, metric) correlate with study sample sizes?
3. What is the exact statistical significance of the overall *ES*?
4. How heterogeneous is the *ES* distribution in the post-Maimonides database?
5. What are the primary moderator variables (target emotionality characteristics, selected participants, GMF, LST, etc.)?
6. What is the exact amount of publication bias in the post-Maimonides database?
7. Can the Maimonides results be replicated by an independent group of researchers?
8. Log as much information about participants and methodology as possible; this information will be very valuable for independent reviewers tracking patterns in the data, and it adds confidence to any inferences drawn from the analysis.

Answering the first six of these questions is a task for future meta-analysts of post-Maimonides DESP studies. The last question is a challenge to a parapsychology laboratory to attempt an exact replication of the Maimonides findings. In our view, both are interesting and worthwhile tasks for contemporary parapsychologists.

Comments on Meta-Analytic Methods

In this section we comment on limitations of meta-analytic methods commonly used in parapsychology meta-analyses and make suggestions for how these methods can be updated and improved.

Effect Size and Significance Level Estimates

The methods commonly used to estimate effect sizes and significance levels in parapsychological meta-analyses of ESP research have almost always made use of fixed-effects models, or models that assume each study is measuring a common *ES* for a common underlying population. Examples of such fixed-effects models are mean effect sizes (unweighted or weighted), Stouffer's *Z* (unweighted or weighted), and exact binomial tests on overall hit rates from trials combined across studies. Although these methods are certainly reliable for testing for the presence of an effect in a meta-analytic database of studies—whether or not the studies have a homogeneous or heterogeneous distribution of effect sizes—they cannot in general be regarded as reliable methods for determining the true *ES* or true hit rate in an underlying population from which the studies in the database draw, particularly when the studies produce a heterogeneous *ES* distribution. The reason for this is simple to understand; when a database of studies is heterogeneous, the study *ES*s come from different populations or are being moderated by significantly different methodological variables in the studies themselves, or both. Consequently, any combined *ES* based on a fixed-effects analysis will be an arbitrary composite of studies with these significantly different characteristics. As a concrete example, consider the STDR post-PRL ganzfeld database, from which we observed that studies using selected individuals produce significantly different hit rates than studies using unselected ones. The overall hit rate of 32.2% is clearly an arbitrary composite of studies using selected and unselected volunteers, and could easily have been significantly higher or lower, depending on the proportion of included studies using selected versus unselected individuals, and vice versa.

What this observation entails is that if we want a reliable estimate of the overall *ES* or hit rate in a heterogeneous database of ESP studies, we must use in place of fixed-effects models, random effects and mixed-effects models[20] (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cochrane, 2002). Whereas fixed-effects models only take into account within-study variances, these latter models take into account between-study variances (due to whatever source). As a result, the larger the heterogeneity in a given database of studies (i.e., the larger the between-study variances), the wider the confidence intervals that random and mixed-effects models will place on the *ES* estimate, in comparison to fixed-effects models. (Of course, when heterogeneity is little to nonexistent in a database, the confidence intervals of random and mixed-effects models will be nearly identical to those of fixed-effects models.) This is what we need when attempting to estimate the overall *ES* in a heterogeneous database of studies. Both random and mixed-effects models could be applied to all of the meta-analytic databases studied in this paper. Towards this end, a meta-analysis of the post-PRL ganzfeld studies that makes use of the Generalized Linear Mixed Model for the binomial distribution (SAS Institute, 2011) is currently underway.

[20] "Mixed-effects" refers to a statistical model that uses both fixed effects and random effects.

Often, though, it is of limited use to know just that there is a significant effect in a heterogeneous database, or just to have an overall *ES* estimate that takes into account between-study variance. We also want to know the sources of the heterogeneity such as the moderator variables. They allow us to block studies into subgroups with less heterogeneity than the original database, creating sometimes entirely homogeneous subgroups. To identify the conditions under which we can reliably expect a study to be replicated successfully in terms of effect size and a pre-specified alpha, we need moderator variables. We saw this with the ganzfeld studies, as evidenced by how studies using selected versus unselected individuals strongly moderated the heterogeneity in the STDR's post-PRL database. Given that selected individuals studies in the forced-choice ESP databases of HF and STDR also produced significantly different overall effect sizes, we can reasonably expect a similar moderation of heterogeneity in those respective databases (This analysis has yet to be carried out, and would be appropriate for a future forced-choice meta-analysis). In doing a moderator variable analysis, it is an advantage that the moderator variables be predicted before analyzing the data to avoid post hoc data analysis concerns. It is also important that not too many moderator variables are predicted at once, as this increases the chance of finding a spurious moderator variable for a given set of studies.

The methods discussed here for estimating *ES*s and significance levels in meta-analytic databases are the most widely used in contemporary meta-analyses in medicine and other fields in behavioral psychology (Borenstein et al., 2009; Cochrane, 2002), and are considered the "state-of-the-art" in meta-analytic methods. In parapsychology, only three meta-analyses have used these methods, namely, the meta-analysis of anticipatory response studies by Mossbridge et al. (2012), the ganzfeld meta-analysis by Tressoldi (2011), and the distant intentions meta-analysis by Schmidt (2012). In keeping with parapsychology's tradition of staying at the forefront of methodological practices, we believe it is time for all future parapsychological meta-analysts to adopt these methods as well. An excellent, self-contained introductory text to these methods is the one by Borenstein et al. (2009).

Heterogeneity Tests

Although we have talked about heterogeneity, we have not discussed methods of testing for it. This is particularly important for parapsychology, as the methods most commonly used for testing for heterogeneity in its meta-analyses are not the most advanced or accurate methods. A common practice is to test for heterogeneity with Box-and-Whiskers and Stem-and-Leaf plots to identify study outliers (Storm et al., 2010; Storm et al., 2012). Homogeneous databases are then formed by removing study outliers (or in the case of HF, using the "10% trim"). However, there are limitations to this approach. First, it is a rough, qualitative assessment of heterogeneity, and heterogeneity can still be present and extreme in a database when tested for by more accurate, quantitative methods such as the chi-square test[21]. Second, although it may tell us that heterogeneity is present in a database, it does not tell us how much heterogeneity is present. Thus, what we really would like are test statistics that tell us, as accurately as possible, when heterogeneity is present or not in a database, and the amount of heterogeneity that is present.

[21] For example, Derakhshani (2014) has observed this in applying the chi-square test to the 102 study 'homogeneous' ganzfeld database formed by STDR by the method of outlier removal.

The chi-square test helps solve the first problem of giving a relatively accurate test of the presence of heterogeneity in a database[22]. We say “relatively accurate” because it has limitations (Borenstein et al., 2009; Cochrane, 2002). Namely, it has low power for meta-analyses composed of a small number of studies or studies with small sample sizes (the latter being very common in meta-analyses of, say, ganzfeld studies). This means that although a statistically significant result may indicate heterogeneity, a nonsignificant result should not be taken as evidence of no heterogeneity. This is also why a *p*-value of .10, rather than the conventional .05, is sometimes used to determine statistical significance. A further problem with the test is that when there are many studies in a meta-analysis, the test has high power to detect a small amount of heterogeneity that may be unimportant in practice.

Some researchers argue that because methodological diversity always occur in a meta-analysis statistical heterogeneity is inevitable (Borenstein et al., 2009; Higgins 2003). Thus, heterogeneity will always exist whether or not we happen to be able to detect it using a statistical test. To address this, methods have been developed for quantifying the amount of heterogeneity across studies, rather than just testing whether heterogeneity is present. A statistic that is widely used in state-of-the-art meta-analyses for quantifying heterogeneity is

$$I^2 = (Q\text{-}df/Q) \times 100\% \qquad (2)$$

Here, *Q* is the particular chi-square statistic known as Cochran’s *Q*, which is identical to Eq. (1), and *df* is the degrees of freedom of the number of studies (Borenstein et al., 2009; Higgins, 2003). The I^2 statistic tells us the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error. Although there are no definite thresholds for interpreting I^2, a rough guide that has been suggested (Borenstein et al., 2009; Higgins, 2003) and is still often used is as follows: 0% to 40% might be negligible, 30% to 60% may represent moderate heterogeneity, 50% to 90% may represent substantial heterogeneity, 75% to 100% may represent considerable heterogeneity. Important also is that the obtained value of I^2 depends on (a) the magnitude and direction of effects, and (b) the strength of the evidence for heterogeneity (e.g., the *p*-value from the chi-square test or a confidence interval for I^2). Also as a rough guide, an I^2 statistic that falls in the range of 0–25% may justify the use of a fixed-effects model for computing *ES* and significance levels. Anything higher would require the use of a random- or mixed-effects model. In general, our view is that even if heterogeneity is relatively low, a meta-analysis should still report both random/mixed- and fixed-effects estimates. Various examples of how to use I^2 in actual meta-analyses can be found in Borenstein et al. (2009). The only parapsychological meta-analyses to use the I^2statistics are the ones by Mossbridge et al. (2012) and Schmidt (2012).

In sum, to make meta-analytic assessments of heterogeneity as accurate and reliable as possible, we believe that all future meta-analyses in parapsychology should use the I^2 statistic in tandem with Cochran’s *Q*, with the guidelines described here and in the accompanying references.

[22] This is why Derakhshani’s (2014) analyses of heterogeneity in the ganzfeld were based on the chi-square test instead of Box-and-Whiskers and Stem-and-Leaf plots.

Publication Bias

As we have seen, the typical approach to testing for publication bias or a file drawer in a meta-analysis is through the use of fail-safe Ns such as the Rosenthal's or the Darlington-Hayes statistic. Although we see nothing inherently flawed in the use of these statistics—and in fact we have argued at length here that they have been valid and reliable as applied to the meta-analytic databases reviewed in this paper—it should be emphasized that they are not the most widely used in contemporary state-of-the-art meta-analyses. Nor are they the most versatile and effective methods for assessing publication bias. More commonly used methods these days are trim-and-fill and Orwin's fail-safe *N* (Borenstein et al., 2009). The former gives a visual and quantitative estimate of the number of missing studies (as well as their *ES*s) that would have to exist to make an asymmetric funnel plot of study effect sizes symmetric again. The latter is a statistic that indicates how many studies with nonsignificant *ES*s (perhaps zero, perhaps greater than zero) would be needed to bring the overall *ES* in a meta-analytic database to a nonsignificant overall value; moreover, this nonsignificant value can be selected by the researcher. So researchers can determine how many studies with a chosen *ES* level would be needed to bring the overall *ES* down to a chosen value, thereby giving more flexibility of analysis than either the Rosenthal or the Darlington-Hayes statistic. Examples of how to use trim-and-fill and Orwin's fail-safe *N* can be found in Borenstein et al. (2009).

We therefore recommend that all future parapsychological meta-analyses make use of the trim-and-fill method, along with Orwin's fail-safe *N*. The Rosenthal and Darlington-Hayes statistics could also be included, but as supplementary rather than primary analyses.

Of course, we not only want to assess publication bias in a meta-analysis, we want to prevent it as much as possible insofar as it can potentially undermine any statistically significant results found in a meta-analysis. To do this, we recommend that all parapsychological researchers preregister their studies in a trial registry, whether or not they anticipate that their studies will eventually be included in a meta-analysis. Two available registries are the Open Science Framework (2014) and the KPU Registry (2014). Having parapsychology journals and funding sources mandate this practice for any submitted publications or research projects would also go a long way towards making it widespread throughout the field.

Experimenter Fraud and Data Manipulation

Concerns about experimenter fraud and data manipulation have always existed within parapsychology, and the history of identified fraud and manipulation is well-known (Kennedy, 2014). We recognize that there is no evidence of fraud or data manipulation in any of the contemporary research paradigms of parapsychology. Nevertheless, to researchers outside the field, the concern is always present whether there is evidence for it or not. And of course, any studies that achieved significant results through experimenter fraud or data manipulation can potentially undermine positive results obtained in a meta-analysis. In our view, the best way for parapsychologists to minimize these concerns is to adopt the most sophisticated practices possible to prevent the possibility of fraud or data manipulation. A good example of this in the history of parapsychology was the methodology of ganzfeld studies conducted at the now defunct PRL (Honorton et al., 1990) and the KPU (Dalton et al., 1996). These studies involved the use of multiple experimenter teams, automated experimental procedures and data recording, and various security measures to prevent cheating by experimenters and/or participants. For whatever reason, these practices have not yet become the standard of parapsychology research. Moreover, other helpful practices exist that have yet to be widespread. Here we concur with the recommendations of Kennedy (2014), who writes: "Parapsychological research organizations and funding sources should require prospective registration of studies, multiple experimenter designs, and data sharing. These requirements for confirmatory experiments would provide the greatest return on investment in research. Parapsychological journals should strongly promote these practices. In addition, a methodologically-oriented colleague can be invited to observe or audit an experiment. These practices should also be standard study quality rating factors in meta-analyses.

Research data in parapsychology should be collected, managed, and analyzed with the expectation that the data will have detailed, critical scrutiny by others. The optimal scientific approach is to make all or part of the raw data openly available. However, when biased post hoc analyses are likely, an original investigator may reasonably require that the recipient register the planned analyses publicly, including corrections for multiple analyses, prior to receiving copies of the data.

An appropriate working assumption is that an experimenter who has competently conducted a study and is confident of the results will readily provide the data for independent analyses. If an experimenter is unwilling to provide the data for independent analyses, an appropriate assumption is that the experiment has questionable methodology and the experimenter has something to hide" (*p*. 11).

In other words, we agree with Kennedy that the above should be the new code of conduct in parapsychology research, and that doing so will not only further strengthen the overall quality and reliability of the evidence (assuming it still holds up or improves under these conditions) but also significantly help in alleviating internal and external concerns relating to the possibilities of experimenter fraud and data manipulation.

A Note on Bayesian Methods

The use of Bayesian meta-analytic methods by skeptics and proponents has become much more common in recent years, as evidenced by the recent analyses of Bem's precognition studies (Wagenmakers et al., 2011; Bem et al., 2011; Rouder and Morey, 2011) and ganzfeld/RV studies (Utts et al., 2010; Tressoldi, 2011; Tressoldi, 2013; Rouder et al., 2013; Storm et al., 2013). This merits some comments, even though the analyses used in this paper are based entirely on classical statistics.

It is noteworthy that when Bayesian meta-analytic methods have been applied to the ganzfeld or remote viewing paradigms, the results (in terms of the computed Bayes factors) consistently and overwhelmingly support the ESP hypothesis against the null hypothesis (Utts et al., 2010; Tressoldi, 2011; Rouder et al., 2013; Storm et al., 2013). However, one disadvantage of Bayesian methods is that they allow for skeptics to arbitrarily choose prior odds and prior distributions so extremely against the ESP hypothesis that no amount of data produced by a parapsychological experiment will be sufficient to flip their priors to posteriors that support the ESP hypothesis. Another disadvantage is that it is rarely clear what necessarily justifies one choice of a prior over another, even though they may differ by several orders of magnitude (particularly in the case of prior odds).

To account for these disadvantages, we would advise that Bayesian methods, to the extent that they are used at all for a meta-analysis, should always be used in parallel with classical statistical methods (such as the ones described in this section). Also essential is that, during the preregistration process, the priors (i.e. both the prior odds and the prior distributions) of the researchers should be listed before the meta-analysis is conducted, along with a detailed document explaining the reasoning that went into the chosen priors. It may also be advisable to allow time for the chosen priors to be debated in a public online forum before the meta-analysis is conducted. This would help ensure that the chosen priors are regarded as reasonable by colleagues, or to reveal ahead of time that they are not regarded as reasonable. It would also help minimize the chances that priors are disputed post hoc, or that different priors are chosen in a subsequent meta-analysis without explanation. These recommendations apply to both skeptics and proponents. For each group, they would help prevent unconscious or conscious shifting of the evidential goalposts—something that Bayesian methods can be easily abused for when subjective biases for or against the possibility of ESP are present.

Discussion

To summarize, we have found that every ESP research paradigm reviewed in this paper has produced cumulative *ES*s and significance levels well above chance expectation. We have also found that each research paradigm has produced results that (in our view) merit further process-oriented and proof-oriented research by both parapsychologists and skeptical but open-minded scientists.

From the analyzed results, the ganzfeld paradigm seems to be the most promising in terms of providing a reliable recipe for producing medium to large effect sizes in tandem with replication rates of 80% or greater; this recipe involves the use of selected participants and other identified moderator variables. The forced-choice paradigm seems the most promising in potentially yielding extremely high replicability rates (90% or greater) with small *ES*s, through the use of selected volunteers and optimal study designs (i.e. using selected subjects and trial-by-trial feedback).

The remote viewing paradigm seems to have produced consistently medium effect sizes over time, even with improvements in study methodological quality; and the data suggests that enhanced effect sizes can be best achieved through participant training in RV techniques, target images with high information entropy, and free reporting of perceptions.

The DESP paradigm in the post-Maimonides era has clearly produced overall effect sizes well above chance expectation, but less than those found in the more successful Maimonides program; the discrepancies in methodology between pre and post Maimonides seem plausible as explanations for this difference under the ESP hypothesis, but little seems to be definitively understood about the key moderator variables for these studies. This would seem to merit more research by parapsychologists into finding out what the key moderator variables are, both through a formal meta-analysis of the existing post-Maimonides database and new D*ES*P studies that closely replicate the Maimonides methodology.

These findings notwithstanding, we believe it is essential for parapsychologists to update their methods of conducting meta-analyses and studies. This update will allow them to confirm or falsify the predictive validity of the most promising results in each paradigm to a degree of certainty not possible with previous methods. In terms of meta-analysis, we believe it is essential that all future meta-analyses (whether of new studies or previous databases) follow the guidelines prescribed in Section 6. In terms of conducting studies, we believe that the practices of preregistration, multiple experimenter designs, and data sharing are absolutely essential to enhancing the actual and perceived reliability of parapsychological data, particularly in the eyes of skeptical but open-minded scientists in the mainstream scientific community.

If, after following our prescriptions, parapsychologists confirm our predicted boosts in replication rates and *ES*s, we believe that this is likely to motivate open-minded scientists to attempt independent replications under the same conditions. Were such scientists successful at replicating the results of parapsychologists, it would mark a decisive turning point in how the evidence for ESP is perceived by the rest of the scientific community. Alternatively, if parapsychologists and/or skeptical but open-minded scientists are unable to confirm our predictions, we will have learned either that ESP (if it exists) likely does not obey lawful statistical patterns (in which case, this would seem to cast doubt on the possibility of ever developing a controlled understanding of it or for that matter convincing the rest of the scientific community of its reality), or that we need a far better understanding of and control over experimenter effects, or that the seemingly promising results of previous ESP studies and meta-analyses were simply spurious. Any of these outcomes, in our view, would constitute significant progress from the present situation.

References

Baptista, J., & Derakhshani, M. (2014). Beyond the Coin Toss: Examining Wiseman's Criticisms of Parapsychology. *Journal of Parapsychology, 78*(1), 56-79.

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

Bem, D. J., Palmer, J., & Broughton, R. S. (2001). Updating the ganzfeld database: A victim of its own success? *Journal of Parapsychology*, 65, 207–218.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data?. *Journal of Personality and Social Psychology*, 101, 716–719. Retrieved from http://dl.dropboxusercontent.com/u/8290411/ResponsetoWag

Bierman, D. & Rabeyron, T. (2013). Can PSI research sponsor itself? Simulations and results of an automated ARV-casino experiments. Proceedings of the 56th Annual Convention of the Parapsychological Association (pp.15).

Borenstein, M., Hedges, L. V., Higgins, J. *P*. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. New York: Wiley. Retrieved from http://onlinelibrary.wiley.com/book/10.1002

Bösch, H. (2004). Reanalyzing a meta-analysis on extra-sensory perception dating from 1940, the first comprehensive meta-analysis in the history of science. In S. Schmidt (Ed.), Proceedings of the 47th Annual Convention of the Parapsychological Association, University of Vienna (pp. 1–13)

Broughton, R. S., Kanthamani, H., & Khilji, A. (1989). Assessing the PRL success model on an independent ganzfeld database. *Proceedings of Presented Papers: The Parapsychological Association 32nd Annual Convention*, 26–33.

Carter, C. (2010). Heads I lose, tails you win, or, how Richard Wiseman nullifies positive results and what to do about it. *Journal of the Society for Psychical Research*, 74. Retrieved from http://www.sheldrake.org/D&C/controversies/Carter_Wiseman.pdf

Child, I.L. (1985). Psychology and anomalous observations: The question of ESP in dreams. *American Psychologist*, 40, 1219–1230.

Child, I.L., Kanthamani, H., & Sweeney, V.M. (1977). A simplified experiment in dream telepathy [abstract], In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in parapsychology 1976 (pp. 91–93). Metuchen, NJ: Scarecrow Press.

Colyer, & Morris, R. (2001). Imagery in the ganzfeld: The effects of extroversion and judging instructions upon imagery characteristics.

Dalton, K., Delanoy, D., Morris, R. L., Radin, D. I., Taylor, R., & Wiseman, R. (1996). Security measures in an automated ganzfeld system. *Journal of Parapsychology*, 60, 129-147.

Dalton, K., Steinkamp, F., & Sherwood, S. .J. (1999). A dream G*ES*P experiment using dynamic targets and consensus vote. Journal of the American Society for Psychical Research, 93, 145–166.

Dalton, K., Utts, J., Novotny, G., Sickafoose, L., Burrone, J., & Phillips, C. (2000). Dream G*ES*P and consensus vote: A replication. Proceedings of the 43rd Annual Convention of the Parapsychological Association, Freiburg, Germany, pp. 74–85.

Darlington, R. B., & Hayes, A. F. (2000). Combining independent *p* values: Extensions of the Stouffer and binomial methods. *Psychological Methods*, 5, 496–515. doi: 10.1037//1082-989X.5.4.496

Derakhshani, M. (2014). *On the statistical replicability of ganzfeld studies.* Manuscript in preparation

Dunne, B.J. & Jahn, R. G. (2003). Information and Uncertainty in Remote Perception Research. *Journal of Scientific Exploration*, 17, 207–241.

Hansen, G. P.; Utts, J. and Markwick, B. (1991). Statistical And Methodological Problems Of The PEAR Remote Viewing Experiments. *Research in Parapsychology 1991*, pp. 103-105.

Harris, M., & Rosenthal, R. (1988). Enhancing human performance: Background papers, issues of theory and methodology. Retrieved from http://www.nap.edu/openbook.php?record_id=779

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128. doi:10.3102/10769986006002107

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*,327, 57-560. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC192859

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

Honorton, C., & Ferrari, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935–1987. *Journal of Parapsychology*, 53, 281–308.

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., 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.

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

Hyman, R., & Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology*, 50, 350–364.

Kanthamani, H., & Broughton, R.S. (1992). An experiment in ganzfeld and dreams: A further confirmation. Proceedings of the 35th Annual Convention of the Parapsychological Association, Las Vegas, NV, pp. 59–73.

Kanthamani, H., & Khilji, A. (1990). An experiment in ganzfeld and dreams: A confirmatory study. Proceedings of the 33rd Annual Convention of the Parapsychological Association, Chevy Chase, MD, pp. 126–137.

Kennedy, J. E. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *Journal of Parapsychology*, 67, 53–74. http://jeksite.org/psi/jp03.htm

Kennedy, J. E. (2014). Experimenter misconduct in parapsychology: Analysis manipulation and fraud. Retrieved from http://jeksite.org/psi/misconduct.htm and http://jeksite.org/psi/misconduct.pdf

KPU Registry for Parapsychological Experiments (2014). Retrieved from http://www.koestler-parapsychology.psy.ed.ac.uk/TrialRegistry.html

Krippner, S. C., & Friedman, H. L. (2010). Editor's Epilogue: Is It Time for a Détente? In Krippner, S., & Friedman, H. L. (Eds.). (2010). Debating Psychic Experience: Human Potential Or Human Illusion?. ABC-CLIO.

Krippner, S., & Persinger, M. (1996). Evidence for enhanced congruence between dreams and distant target material during periods of decreased geomagnetic activity. *Journal of Scientific Exploration*, 10, 487–493.

Krippner, S., Vaughan, A., & Spottiswoode, S.J.*P*. (2000). Geomagnetic factors in subjective precognitive dream experiences. *Journal of the Society for Psychical Research*, 64, 109–117.

May, E. C. (2007). Advances in anomalous cognition analysis: A judge-free and accurate confidence-calling technique. Proceedings of the 50th Annual Convention of the Parapsychological Association (pp.57–63).

May, E. C., Marwaha, S. B., & Chaganti, V. (2011). Anomalous cognition: two protocols for data collection and analyses. *Journal of the Society for Psychical Research*, 75, 191-210.

May, E., C. (2011). Possible thermodynamic limits to anomalous cognition: Entropy gradients. *Journal of the Society for Psychical Research*, 75, 65-75.

McDonough, B. E., Don, N. S., & Warren, C. A. (1994). EEG in a ganzfeld psi task. In D. J. Bierman (Ed.), *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention*, Durham, North Carolina, 273-283.

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

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., 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.

Morris, R., Dalton, K., Delanoy, D. & Watt, C. (1995). Comparison of the sender/ no sender condition in the ganzfeld. *Proceedings of presented papers: The Parapsychological Association 38th Annual Convention*, Durham, North Carolina, 244-259

Morris, R., Summers, J., & Yim, S. (2003). Evidence of anomalous informa-tion transfer with a creative population. Proceedings of the Parapsychological Association 46th Annual Convention (pp. 116–131).

Mossbridge, J., Tressoldi, P., & Utts, J. (2012). Predictive physiological anticipation preceding seemingly unpredictable stimuli: A meta-analysis. *Frontiers of Psychology*, 3, 390. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fpsyg.2012.00390/abstract

Mosseau, M-C. (2003). Parapsychology: Science or pseudoscience? *Journal of Scientific Exploration*, 17, 271–278. Retrieved from http://www.scientificexploration.org/journal/jse_17_

Open Science Framework (2014). Retrieved from https://osf.io/getting-started/

Parra, A., & Villanueva, J. (2006). ESP under the ganzfeld, in contrast with the induction of relaxation as a psi-conducive state. *Australian Journal of Parapsychology*, 6, 167–185.

Persinger, M.A., & Krippner, S. (1989). Dream ESP experiments and geomagnetic activity. *Journal of the American Society for Psychical Research*, 83, 101–116.

Roe, C. A., & Flint, S. (2007). A remote viewing pilot study using a ganzfeld induction procedure. *Journal of the Society for Psychical Research,* 71, 230–234.

Roe, C. A., Hodrien, A., & Kirkwood, L. (2012). Comparing remote viewing and ganzfeld conditions in a precognition task. Abstracts of presented papers: Parapsychological Association 55th Annual Convention, pp. 36-37.

Roe, C., Cooper C. & Martin, H. (2010). A comparison between remote viewing and ganzfeld conditions in a precognition task. Proceedings of the 53rd PA Conference, pp.21-22.

Rosenthal, R. (1986). Meta-analytic procedures for social science research. *Educational Researcher*, 15, 18. doi: 10.2307/1175262

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem’s ESP claim. *Psychonomic Bulletin & Review*, 18, 682– 689. doi: 10.3758/s13423-011-0088-7

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes factor meta-analysis of recent extrasensory perception experiments: Comment on Storm, Tressoldi, and DiRisio (2010). *Psychological Bulletin*, 139, 241–247. doi:10.1037/a0029008. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23294092

Ryan, A. (2009). Geomagnetic activity & dreams. Proceedings of the 33rd International Conference of the Society for Psychical Research, Winchester, U.K. Retrieved from http://www.greyheron1.plus.com/

SAS Institute (2011). SAS/STAT 9.3 User's guide: The GLIMMIX procedure. Cary, NC: SAS Institute.

Saunders, D. R. (1985). On Hyman's factor analysis. *Journal of Parapsychology*, 49, 6-88.

Scargle, J. D. (2000). The file-drawer Problem in scientific inference. *Journal of Scientific Exploration*, 14, 91–106.

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.

Schmidt, S. (2012). Can we help just by good intentions? A meta-analysis of experiments on distant intention effects. *Journal of Alternative and Complementary Medicine*, 18, 529-533. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22784339

Schwartz S. (2014). Through time and space: The evidence for remote viewing. In D. Broderick & B. Groetzel (Eds.), The evidence for psi. Jefferson, NC: McFarland.

Sherwood, S. J., & Roe, C. A. (2003). A review of dream ESP studies conducted since the Maimonides dream ESP programme. *Journal of Consciousness Studies*, 10(6,7), 85-109.

Sherwood, S.J., Dalton, K., Steinkamp, F., & Watt, C. (2000). Dream clairvoyance study II using dynamic video-clips: Investigation of consensus voting judging procedures and target emotionality. *Dreaming*, 10, 221–236.

Spottiswoode, S.J.P. (1997). Geomagnetic fluctuations and free-response anomalous cognition: A new understanding. *Journal of Parapsychology*, 61, 3–12.

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. (2003). Remote viewing by committee: RV using a multiple agent/multiple percipient design. *Journal of Parapsychology*, 67, 325–342.

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.

Storm, L., Tressoldi, P. E., & DiRisio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, 136, 471-485. doi: 10.1037/a0019457

Storm, L., Tressoldi, P. E., & DiRisio, L. (2012). Meta-analysis of ESP studies, 1987–2010: Assessing the success of the forced-choice design in parapsychology. *Journal of Parapsychology*, 76, 242.

Storm, L., Tressoldi, P. E., & Utts, J. (2013). Testing the Storm et al. (2010) meta-analysis using Bayesian and frequentist approaches: Reply to Rouder et al. (2013). *Psychological Bulletin*, 139, 248–254. doi:10.1037/a0029506. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23294093

Subbotsky, E. & Ryan, A. (2009). Motivation and belief in the paranormal in a remote viewing task. Retrieved from https://www.researchgate.net/publication/236984643_Motivation

Symmons, C., & Morris, R. (1997). Drumming at seven Hz and automated ganzfeld performance. The Parapsychological Association 40th Annual Convention: Proceedings of Presented Papers (pp. 441-454)

Targ, R. (1994). Remote viewing replication evaluated by concept analysis. *Journal of Parapsychology*, 58, 271-284.

Targ, R., & Katra, J. E. (2000). Remote viewing in a group setting. *Journal of Scientific Exploration*, 14, 107–114.

The Cochrane Collaboration (2002). Retrieved from http://www.cochrane-net.org/openlearning/html/mod13-4.htm

Tressoldi, P. E. (2011). Extraordinary claims require extraordinary evidence: The case of non-local perception, a classical and Bayesian review of evidences. *Frontiers of Psychology*, 2, 117. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fpsyg.2011.00117/abstract

Tressoldi, P. E. (2013). How much evidence is necessary to define the reality of a phenomenon? Frequentist, Bayesian, and quantum modeling of ganzfeld ESP. In Advances in parapsychological research 9. Jefferson, NC: McFarland.

Ullman, M., Krippner, S., with Vaughan, A. (2003). Dream telepathy: Experiments in nocturnal ESP. Hampton Roads.

Utts, J. (1996). An assessment of the evidence for psychic functioning. *Journal of Scientific Exploration*, 10, 3-30.

Utts, J., Norris, M., Suess, E, & Johnson, W. (2010, July). The strength of evidence versus the power of belief: Are we all Bayesians? Paper presented at the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia. Retrieved from https://www.stat.auckland.ac.nz/en.html

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology,* 100, 426–432. Retrieved from http://www.ejwagenmakers.com/2011/WagenmakersEtAl2011_JPSP.pdf

Watt, C. & Nagtegaal, M. (2004). Reporting of blind methods: An interdisciplinary survey. *Journal of the Society for Psychical Research*, 68, 105–114.

Watt, C. (2006). Research assistants or budding scientists? A review of 96 undergraduate student projects at the Koestler Parapsychology Unit. *Journal of Parapsychology*, 70, 355-356. doi: Retrieved from http://www.koestlerparapsychology.psy.ed.ac.uk/cwatt/Documents/WattJ

Watt, C. (2014). Precognitive dreaming: Investigating anomalous cognition and psychological factors. Journal of Parapsychology, 78(1), 115-125.

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