Journal of Parapsychology, 78(1), 56–79

BEYOND THE COIN TOSS: EXAMINING WISEMAN'S CRITICISMS OF PARAPSYCHOLOGY

By Johann Baptista and Max Derakhshani*

ABSTRACT: We examine the critique of parapsychology offered by Professor Richard Wiseman in his 2010 paper, *Heads I Win, Tails You Lose; How Parapsychologists Nullify Null Results*, published in the *Skeptical Inquirer*, and offer detailed rebuttals to his main contentions. Some of the analyses we conduct are as follows: We compare reproducibility of psi experiments to reproducibility of experiments across related mainstream fields, finding that they are similar. Using both theoretical and empirical approaches, we demonstrate that file-drawer effects are not significant in the ganzfeld. We scrutinize and critique cases of alleged experimenter nullification of null results. We challenge—and offer alternatives to—the conclusions of the Milton and Wiseman meta-analysis, based on findings from Bem, Palmer, and Broughton, as well as our own results. We show that the evidence for ostensible declines in the actual effects of ganzfeld and forced-choice ESP paradigms is largely illusory and challenged by findings of recent inclines. Finally, we present strategies for progress according to the most compelling trends and consistencies we have found in the present database. These results, we hope, serve an illustrative purpose: a case examination of criticism in parapsychology with Wiseman as the main example, showing the degree to which the literature seems to support psi as the most plausible explanation of the data.

Keywords: Wiseman, ganzfeld, critique, skepticism, psi, parapsychology

Written in the spirit of the contributions made to Krippner and Friedman's (2010) book, *Debating Psychic Experienc*e, we aim in this essay to contribute to the ongoing conversation on psi and science. Many reviews and meta-analyses have been published which examine the data, including very recent ones—our aim is to examine the criticism. For this purpose, we selected a well-known general critique of the field by Wiseman (2010a).

The arguments of that critique have not been extensively rebutted before. Carter (2010a), in "*Heads I Lose, Tails You Win: How Richard Wiseman Nullifies Null Results and What To Do About It,*" replied to Wiseman, but his rejoinder concentrated most heavily on Wiseman's own conduct as an experimenter and not so much on his arguments. We address the latter to the best of our ability, and we keep our analysis manageable by placing special emphasis on the ganzfeld experiments, the "flagship" of parapsychology (Parker, 2000).

In the interest of full disclosure, our position is that these experiments and others have produced robust evidence for a communications anomaly of the type outlined by Bem and Honorton (1994)—though we reserve opinion on whether this is, ipso facto, psi—to such a degree that they necessitate analysis and replication from the mainstream. This is due both to careful precautions of investigators over the years as well as to surprising consistencies in the data, which we explore. Our paper ends with a point of agreement between us and Wiseman, illustrating the possibilities for future research.

If our comments and suggestions aid the development of parapsychology as a field, or conversely, the improvement of skeptical analysis, we will consider our job well done.

The Perception of Null Results

The major premise of Wiseman's critique is that parapsychologists tend to accept positive results as evidence for psi but dismiss null results with post hoc explanations. In this regard, Wiseman writes:

> Parapsychologists frequently create and test new experimental procedures in an attempt to produce laboratory evidence for psi. Most of these studies do not yield significant results. However, rather than being seen

> as evidence against the existence of psychic ability, such null findings are usually attributed to the experiment being carried out under conditions that are not psi-conducive. (Wiseman, 2010a, p. 37)

Crucial to the strength of Wiseman's critique is the question of how much weight null results should reasonably carry in the assessment of the evidence for psi—and what kind of null results are at issue. But before we address this, we note that although it is true that most studies in parapsychology databases do not display significant results, it is also true that the number that *do* is significantly above the null hypothesis expectation. Consider, for example, the post-PRL database, which consists of the studies in the Milton and Wiseman (1999) and Storm, Tressoldi, and Di Risio (2010) meta-analyses, covering the period 1988–2008. These 60 studies were conducted following a seminal report from Honorton's Psychophysical Research Laboratories (PRL; Bem & Honorton, 1994), after the strict methodological guidelines proposed by Hyman and Honorton (1986). Only 15 of these post-PRL studies (25%) were significant at $p \leq .05$, whereas under the null, only 5% should have met this threshold, and the probability of getting 15 or more significant studies by chance alone is less than 1 in 5,200,000. Thus, average investigators have a probability of producing significant results that is five times what they would have if nothing significant was occurring in these experiments. We consider that important. Indeed, it is on *this* sort of observation that the ganzfeld, and similar domains of research, rest their claim to repeatability.

But is it sufficient? We note that there are several valid metrics by which to gauge reproducibility, and it is beyond the scope of this paper to present them all (see Cumming, 2012; Utts, 1991). The metric we focus on is the proportion of significant studies ($p \leq .05$) produced by a given research technique, a result governed by statistical power, or 1-β. This can be thought of as the probability of obtaining significance *given the attributes* of one's research methodology, and it is a direct function of type of significance test, effect size (ES), sample size (N), and alpha (α) level. Because power governs the potential success of a study, we believe it critical to consider power before judging what level of reproducibility one *should* be seeing in a field as a litmus test of validity; after all, when power is low, we will fail to detect even a completely consistent effect more often than not.

In this vein, it is reasonable to ask how much power is employed in parapsychology, generally. According to Utts (1991) and Tressoldi (2012), not a lot. Taking the ganzfeld as a prototypical example, for the 105 studies reported in Storm, Tressoldi, and Di Risio (2010) with four-choice designs, the overall hit rate is 32.2% and the mean sample size is 42, for an average power per study of about 30%; this value comes close to the proportion of significant studies (28.5%) in that sample. Similar calculations performed by Derakhshani (2014)—using his own power test and one recommended by Ioannidis and Trikalinos (2007)—demonstrate that the proportion of significant studies in all past ganzfeld databases can be accurately predicted using standard power assumptions, within 95% confidence intervals. This suggests that ganzfeld studies elicit the level of consistency that is expected given the characteristics of those studies, and that they are replicable insofar as we can make predictions about their probability of success and have them verified. The evidence is that psi effects, at least in the ganzfeld, lawfully follow the predictions of conventional statistical models to a degree that is conducive to scientific investigation.

We should thus be able to reliably effect changes in our levels of success, using these models. If we aim for 80% power in the ganzfeld, for example, we may try increasing sample size alone; however, this will result in at least 236 needed trials (given the 32.2% hit rate found in Storm et al., 2010)—a quantity likely to be inaccessible to the average investigator. In fact, the largest number of trials ever run in a single ganzfeld study is 138 (Parra & Villanueva, 2006). Another option to boost power is to raise the ES of studies. Derakhshani (2014) takes this route and shows, based on the post-PRL database, that if investigators use only selected participants (e.g., participants with prior psi experience, mental discipline practice, prior psi training, belief in psi, or preferably a combination of these)—a population that achieved a 40.1% hit rate in the post-PRL database—they would need only 56 trials for 80% power. We note that this predicted higher proportion of significant studies is not only completely consistent with past findings, but practicably attainable.

Another question we might ask about power and replication in parapsychology is how they stack up with what is found in other sciences. To our knowledge, there has never been an in-depth comparison of this type, but one is sorely needed. For example, in Richard, Bond, and Stokes-Zoota's (2003) exhaustive meta-analysis of 322 meta-analyses in social psychology, the average statistical power was 20%, a little below that of the post-PRL database. With this power, the typical social science experiment would need at least 173 trials to achieve 80% reproducibility (at $p \leq .05$), which is already considerably higher than normal (Hartshorne & Schachner, 2012). The reason for this is that ESs in social psychology are usually small—about $r = .21$ on average—and researchers tend not to conduct

large enough studies to compensate for this. In fact, almost a third of the ESs reported in Richard et al. (2003) were $r = .1$ or below, requiring an average N of 772 just to achieve a power of 80% (Hartshorne & Schachner, 2012).

Hartshorne and Schachner (2012) write, additionally, that

> according to multiple meta-analyses, the statistical power of a typical psychology or neuroscience study to detect a medium-sized effect (defined variously as $r = .3$, $r = .4$, or $d = .5$) is approximately .5 or below (Bezeau & Graves, 2001; Cohen, 1962; Kosciulek & Szymanski, 1993; Sedlmeier & Gigerenzer, 1989). (p. 2)

But in fact, for small effects ($d \leq 3$), this power is much lower. Rossi (1990) observed a mean power of 17% across 221 articles for ESs in this range, in three prominent psychology journals starting in 1982. Neuroscience research has also been recently reviewed by Button et al. (2013), who looked at 730 studies in 49 meta-analyses and concluded that the median statistical power for that discipline was about 21%. They subsequently observed that the removal of seven outlying meta-analyses with very large effect sizes brought their power estimate to 18%. All of these power values—from the average power in social psychology, the mean power for small effects in psychology, and the median power for neuroscience studies—fail to meet the average power for a ganzfeld study conservatively calculated at 30%, for all 105 studies in Storm et al. (2010). Considering just the recently gathered 30 ganzfeld studies from 1997 to 2008 (Storm et al., 2010), the average power is actually much higher, at approximately 43%. Even for all the nonganzfeld free-response studies reported during that period in Storm et al. (2010), the mean power of 19% (excluding four studies not of four-choice design) is still marginally greater than for most of the aforementioned mainstream areas.

Bakker, Van Dijik, and Wicherts (2012) estimate, moreover, that for the average ES of psychology research ($d = 0.5$, which they note is skewed by publication bias), using a two independent samples comparison, the power of psychology studies across multiple meta-analyses is about 35% (p. 544). Despite the roughness of this estimate, it happens to closely match the reported current proportion of significant results in the Reproducibility Project database (33.3%), a meta-experiment with a median power of 95% to detect effects across a wide range of replications of papers representatively sampled from psychology journals (Nosek, Lai, LeBel, Gilbert, & Strohminger, 2014). Why are these percentages so similar? The answer is that publication bias in psychology is very prevalent, so that if we assume a simplified model reasonably close to the truth, all published psychology studies are significant. For psychology studies with true effects, then, following Derakhshani (2014) and Ioannidis and Trikalinos (2007), our mean power estimate says that 35% will reach significance and get published. Therefore 35% should *very* roughly be the proportion of significant published studies with true effects. The other 65% should be false positives drawn from studies with no true effects. So when experiments such as those in the Reproducibility Project, using extremely high power, representatively replicate from all published significant studies, those 35% of studies with true effects should be the studies that are successfully replicated. Since this seems to be the case, it confirms the predictions of Derakhshani (2014) and Ioannidis and Trikalinos (2007) that, in the presence of a consistent effect, the average power in a field should serve as a good quantifier of reproducibility, per our definition.

On this subject, Nosek (2012) writes:

> There exists very little evidence to provide reproducibility estimates for scientific fields, though some empirically informed estimates are disquieting (Ioannidis, 2005). When independent researchers tried to replicate dozens of important studies on cancer, women's health, and cardiovascular disease, only 25% of their replication studies confirmed the original result (Prinz, Schlange, & Asadullah, 2011). In a similar investigation, Begley and Ellis (2012) reported a meager 11% replication rate. (p. 657)

In the face of these reproducibility estimates, we argue that for any area of parapsychology to achieve a replication rate of 25% to 30% to 37%—the proportion of significant results in the post-PRL, the whole ganzfeld, and the most recent 30 studies, respectively (Storm et al., 2010)—which we have shown to be comparable to other sciences; is in fact quite remarkable, given that the total human and financial resources devoted to psi research from 1882 to 1993 has been estimated to comprise less than two months' research in conventional psychology (Schouten, 1993, p. 316). This observation warrants the conclusion that not only is the ganzfeld technique consistent, but it is also progressing at a rate similar to that of mainstream social and behavioral fields—and surprisingly so, given its

resources. The conformance of the ganzfeld database to power predictions, moreover, strongly suggests that adoption of strategies to boost power would improve reproducibility, and that attempting to do so would be a worthwhile venture.

Investigating the File Drawer

For a meta-analysis to be valid, arguably the most important criterion is that all of the data are there to analyze—or at least that no systematic bias of any importance is present in the studies selected. Yet this is what Wiseman (2010a) seems to imply by his comments:

> Once in a while one of these [parapsychology] studies produces significant results. Such studies frequently contain potential methodological artifacts, in part because they are using new procedures that have yet to be scrutinized by the research community ... the evidential status of these positive findings is problematic to judge because they have emerged from a mass of nonsignificant studies. Nevertheless, they are more likely than nonsignificant studies to be presented at a conference or published in a journal. (p. 37)

Firstly, it is important to note that the idea that positive studies are more likely to contain methodological artifacts is poorly supported for research into ESP (though it does receive some support for recent research into psychokinesis, as seen below). We are aware of one meta-analysis by Schmidt, Schneider, Utts, and Walach (2004) that found a significant negative correlation between overall quality and ES for direct mental interaction with living systems (DMILS) studies, but not remote staring studies. These correlations are rare. Storm et al. (2010) showed, for example, that for their free response studies conducted from 1992–2008, quality ratings obtained under blind conditions did not correlate significantly with ESs: $r(65) = .08$, $p = .11$. We were further able to demonstrate that for groups of high-scoring selected participants in Storm et al.'s (2010) 30-study ganzfeld database, the mean study quality rating was greater than for the significantly lower-scoring unselected participants ($q = 0.84$ and 0.79, respectively, where $q = 1.00$ was the highest possible rating). We give a sampling of the literature on the question of quality-ES correlations as follows, endeavoring to use only the most recent results for each paradigm of research:

1. The only meta-analysis of physiological presentiment studies conducted to date detected a nonsignificant *positive correlation* between methodological stringency and ES: $r = .21$, 95% CI = –.20–.53 (Mossbridge, Tressoldi, & Utts, 2012).

2. A meta-analysis of forced-choice precognition studies yielded a very small and nonsignificant *positive* correlation between ES and study quality; $r = .08$, $p = .20$, two-tailed (Honorton & Ferrari, 1989).

3. In a review of the success of the forced-choice ESP paradigm in parapsychology, a very small and nonsignificant negative relationship was found between ES and quality ratings, and thus no dependency; $r = -.08$, $p = .48$, two-tailed (Storm, Tressoldi, & Di Risio, 2012).

4. Bösch et al. (2006) found a highly significant correlation between ES and safeguard sum score in their database of RNG studies, indicating that lower quality studies produced larger ESs: $r(386) = .15$, $p = .004$. They noted, however, that the average quality of these studies was very high.

In view of these considerations, the hypothesis that experimental flaws are systematically and inversely related to study ES in parapsychology should be seen as generally unsupported by the evidence, unless analyses using novel quality ratings find conflicting results.

Wiseman's main criticism, however, raises a concern that parapsychologists have been conscious of for decades: the file-drawer problem. Its premise is that studies with positive results are more likely to find their way into meta-analytic databases than studies with negative results, and that this therefore creates a systematically biased sample. This effect has been well-documented (Ahmed, Sutton, & Riley, 2012; Fanelli, 2010; Rothstein, Sutton, & Bornstein, 2005). Fanelli (2010), for example, observed that 84% of publications in various sciences reported positive results—a very unlikely proportion given the low power estimates discussed in the previous section of

this paper—with psychology reporting the most: 91.5%. This estimate for psychology is only minimally different from previous values reported by Sterling (1959) and Sterling, Rosenbaum, and Weinkam (1995), at 97% and 96% respectively. It is common practice for journals to reject null studies in favor of positive ones—the result being that many unsuccessful studies never make it to publication, and thus escape detection by meta-analysts. Even if a study does get into print, it may still be excluded from meta-analytic consideration; biases inherent in the meta-analytic search process or inclusion criteria may cause the study either to be overlooked or disregarded. We make a distinction between these two types of selection bias, calling the first *publication bias* and the second *inclusion bias* (although both are problematic, the former is arguably more so, as unpublished studies are less likely to be found than published studies).

Based on these reasons, then, we note that the selection bias criticism is *a priori* an extremely powerful one, but as we hope to show for parapsychology, ultimately untenable. One reason is that awareness of the filedrawer came early for psi researchers. The earliest systematic cross-laboratories meta-analysis in scientific history, reported in *Extra-Sensory Perception After Sixty Years* (Rhine, Pratt, Stuart, Smith, & Greenwood, 1967), included a statistical method to estimate the influence of publication bias. Additionally, in 1975, the Parapsychological Association (PA) became the first scientific organization to adopt an official policy of publishing null results (Carter, 2010a). Beyond explicitly minimizing the file drawer, this decision brought into common psi research practice techniques designed to measure study selection bias, such as funnel plots, Rosenthal's fail-safe *N*, and trim-and-fill methods, all of which have been used in reviews of psi research to argue effectively against the file-drawer explanation.

With regard to the ganzfeld, for example, Storm et al. (2010) applied Rosenthal's fail-safe *N* (Harris & Rosenthal, 1985, p. 189) and found that no fewer than 2,414 unpublished studies with overall null results (i.e., $z = 0$) would have to exist to reduce their 108 ganzfeld study database to nonsignificance. This is not a likely scenario. However, some have argued that Rosenthal's calculation overestimates the file drawer (Scargle, 2000) by definition, because it implicitly assumes the reservoir of unpublished studies to be unbiased ($z = 0$) instead of directionally negative ($z < 0$). To overcome this problem, there are more conservative procedures such as the Darlington and Hayes (2003) method, which allows for a large proportion of unpublished studies to have negative *z* scores. Applying this method as an additional check for the same homogeneous 102-study database, Storm et al. (2010) showed that the number of unpublished studies necessary to nullify just their 27 studies with statistically significant positive outcomes was 384, and 357 of these could have $z < 0$. Given the official policy of publishing null results set down by the PA, and the small number of scientists conducting research in this area, such a large number of negative studies can only be deemed highly untenable.

With regard to the validity of Rosenthal's fail-safe *N*, we agree with the technical correction put forward by Scargle (2000) that the theoretical mean *z* of unpublished studies for an extreme file-drawer case, under a null distribution, is -0.1085, not 0. Harris and Rosenthal (1988) note, however, that "Based on experience with meta-analyses in other domains of research (e.g., interpersonal expectancy effects) the mean *z* or effect size for nonsignificant studies is not zero but a value pulled strongly from zero toward the mean *z* or mean effect size of the obtained studies (Rosenthal & Rubin, 1978)" (p. 45). Their assumption that the average *z* score of excluded studies is zero is therefore a conservative one for most any distribution that is shifted some positive distance from a null distribution, and although this specifically indicates situations where an effect is present, we argue that the evidence for such an effect in the ESP literature is overwhelming, whatever one may believe about its underlying cause. Another conservative assumption in Rosenthal's procedure is that each excluded study is considered to have a sample size equal to the average sample size of the meta-analysis, whereas overlooked studies tend to be smaller.

Further evidence against the file-drawer effect in the ganzfeld, supporting the notion that unpublished studies show directionally positive results, comes from a mail survey by Blackmore (1980), who queried parapsychologists conducting ganzfeld experiments to obtain a direct estimate of the file drawer. The returned questionnaires revealed 32 unreported studies, 12 of which were still in progress, and one that could not be analyzed. Of the 19 remaining, 14 were judged to have adequate methodology, including 5 that were significant (36% of the total). This proportion of significant results is statistically unlikely according to the null hypothesis; in fact, it yields an exact binomial result of $p = .0004$, or odds against chance of 2,342 to 1. So the file drawer itself is—directly counter to the skeptical prediction—inclined towards the psi hypothesis. Furthermore, the proportion of significant studies in Blackmore's 1980 paper (5 out of 14, or 36%) is not significantly different from the proportion found in Honorton's 1985 meta-analysis (12 out of 28, or 43%), Fisher's exact $p = .46$, one tailed. Given this information, it is not surprising that Blackmore (1980) concluded that "the bias introduced by selective reporting of ESP ganzfeld studies is

not a major contributor to the overall proportion of significant results" (p. 217). Blackmore's survey must be understood in context, however; it took place more than 34 years ago, and 20 studies in it were destined for publication. As such, it can only be considered a snapshot of the file drawer at a given time.

Additionally, even if one entertains the notion that the included ganzfeld studies are drawn from an overall statistically null distribution—in spite of the results of the conservative Darlington-Hayes calculation and the Blackmore (1980) survey—the parapsychological practice of considering significantly negative results to be "psi-missing," and therefore potential evidence for psi, helps to ensure that the negative tail of this distribution is also included, meaning that the average *z* of the excluded studies should be relatively close to zero, not highly negative. This symmetrical exclusion principle is supported by Harris and Rosenthal's (1988) assessment of the ganzfeld, which yielded evidence consistent with "larger positive and larger negative effect sizes than would be reasonable" (Harris & Rosenthal, 1980, p. 44), although by a small margin.

Perhaps most persuasively, as we showed in the first section of this paper, the average power of ganzfeld studies across databases accurately predicts their proportion of significant results, suggesting minimal or no selection bias (Ioannidis & Trikalinos, 2007). Similar calculations to Rosenthal's and Darlington and Hayes', as well as funnel plots and trim and fill algorithms, have plausibly written the file-drawer explanation out of other paradigms in parapsychology, including remote viewing studies (Tressoldi, 2011), psychokinesis studies (Radin et al., 2006), forced-choice ESP studies (Tressoldi, 2011), and precognition studies (Honorton & Ferrari, 1989). Collectively, they provide evidence that selective reporting is not a significant factor in psi research.

There is, however, a still more direct way to tackle Wiseman's (2010a) criticism, since in his words "... only one paper has revealed an insight into the potential scale of [the file-drawer] problem"(p. 37). That paper is the Watt (2007) Koestler Parapsychology Unit report, which surveyed all parapsychology undergraduate projects undertaken and supervised by the Edinburgh staff between 1987 and 2007. About it, Wiseman (2010a) says:

> Only seven of the 38 studies had made it into the public domain, presented as papers at conferences held by the Parapsychological Association ... there was a strong tendency for parapsychologists to make public those studies that had obtained positive findings, with just over 70 percent (five out of seven) of the studies presented at conferences showing an overall significant result, versus just 15% (3 out of 20) of those that remained unreported. (p. 37)

At first glance, this appears to be incontestable proof of a serious publication bias, but a closer look at what Wiseman says is instructive. First, the very fact that meta-analyses in parapsychology include studies presented at conferences but not published in journals (an uncommon practice in the sciences) testifies to its attempt to combat selective reporting (note that PA conference papers are still peer reviewed). Second, Wiseman makes a critical mistake when he mixes projects as varied as "dowsing for a hidden penny, the psychokinetic control of a visual display of a balloon being driven by a fan onto spikes, presentiment of photographs depicting emotional facial expressions, detecting the emotional state of a sender in a telepathy experiment, ganzfeld studies, and card guessing" (p. 37) and then gives the inflated 70% and 15% figures as evidence for a massive file-drawer effect. Because these studies fall into different experimental paradigms, and some of them do not belong clearly to any defined line of research (i.e., they are purely exploratory), mixing them together tells us nothing about the evidential impact of this file drawer on proof-oriented meta-analyses.

It can be seen, for example, that if just one type of study is taken from Edinburgh's varied selection—ganzfeld studies—Wiseman's criticism is rendered moot. Of the 38 KPU undergraduate projects that tested for a psi effect, only 5 were ganzfeld (one by Colyer and Morris, cited by Watt, 2006; one by Morris, Cunningham, McAlpine, and Taylor, 1993; two by Morris, Summers, and Yim, 2003; and one by Symmons and Morris, 1997). Furthermore, although the nonsignificant Colyer and Morris study was the only study not presented at PA conventions, the Morris et al. (1993) study *was* presented, and was also nonsignificant. This leaves a single study in the file drawer whose reasons for not being included are unknown, and whose exclusion is not enough to say anything meaningful about selective reporting in the ganzfeld.

Putting aside ganzfeld studies, three additional student projects were presented at the PA conventions, and they were all DMILS studies. Two had significant results (Brady & Morris, 1997; Delanoy & Sah, 1994) and one was nonsignificant (Watt, Hopkinson, &Fraser, 2006). Examination of Watt (2007) revealed that five of these studies—by Howat et al., Juniper et al., Martin and Miller, Phillips and Morris, and Robert et al.—were not presented at

a PA convention; three of these had nonsignificant but positive results; and two had results that could not be determined from the report. Considering only the three whose results were cited, Watt (2007) writes:

> None was statistically significant; however two were quite low-powered with only 28 participants each, making statistical significance difficult to achieve. All three found effects in the predicted direction and of a magnitude ($r = .15$) larger than that found in the Schmidt, Schneider, Utts, and Walach (2004) meta-analysis of 15 remote staring studies (with the latter finding a mean effect size $d = 0.13$, which converts to $r = .079$). However, the two databases are not independent because Schmidt et al. retrieved unpublished studies, including two of the three studies reported here (Howat, and Juniper & Edlmann). (p.348)

Furthermore, Schmidt et al. (2004) made an effort to locate unpublished experiments; they contacted authors of all the published studies to ask for assistance and posted a search request on an e-mail forum discussing parapsychological research issues. Watt's DMILS studies are therefore unlikely to serve as evidence of a significant file drawer.

The KPU ganzfeld pool, however—because of Watt's comprehensive survey—is an example of a dataset that we can reasonably infer possesses no selective reporting of studies. If we consider the five studies provided, including the Colyer and Morris study, for a total of 195 trials and a hit rate of 33.8%, the cumulative probability of their results under the null hypothesis is $p = .004$ (one-tailed, exact binomial). The 10-study PRL database, too, is known to have no selective reporting; Bem and Honorton (1994) explicitly stated that "the 11 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) (Note: it is common in analyses of the PRL studies that one highly successful study, Study 302, is removed from analysis due to well-known concerns about optional stopping, thereby leaving 10 studies). Additionally, Honorton (1985) 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). This file drawer free database has a hit rate of 32.2%, 329 trials, and a binomial probability of $p = .002$. Given that these hit rates are not significantly different from each other, we can merge the two datasets to form one 15-study pool with no file drawer, 524 trials, a hit rate of 32.8%, and a binomial probability of $p = 5.91 \times 10^{-8}$. This composite hit rate (32.8%) is close to that of the remaining 90 studies in Storm et al.'s (2010) database. When we remove these 15 studies, as well as 3 not of four-choice design, there remain a total of 3,516 trials with a composite hit rate of 31.8%. This convergence of results from three analyzed study pools (KPU, PRL ganzfeld, and the rest of the ganzfeld studies in the Storm et al. database) suggests that if there is a contribution from selective reporting to the overall hit rate, it is likely to be negligible or nonexistent. It is also an example of a surprising consistency in psi research.

In sum, although we acknowledge that we cannot comment as extensively on other paradigms of parapsychology as we can on the ganzfeld, at present we believe that the ganzfeld has performed admirably with regard to the file drawer. If this protocol can be considered representative of parapsychology as a whole, selective reporting of positive results cannot be considered to have significantly influenced the evidence for the existence of psi phenomena.

Parapsychology and Null Results

Wiseman's (2010a) next major criticism involves variations in the procedure of parapsychology experiments:

> If a procedure seems to yield significant psi effects, additional follow-up studies using that procedure are conducted. Although these additional studies occasionally take the form of strict replications, they usually involve some form of variation. If these follow-up studies obtain significant results, they are often the subject of considerable debate: proponents argue that the findings represent evidence of psi, and skeptics scrutinize the work for possible methodological and statistical shortcomings. However, any failure to replicate can be attributed to the procedural modifications rather than to the nonexistence of psi. (p.37)

Although Wiseman's critique in some respects makes a legitimate point, it should be remembered that counter-advocates spend much of their time doing just what Wiseman opposes, but in reverse, and this has been well documented (Carter, 2010a). It is thus important to analyze instances of claimed spurious nullification to determine whether they represent (as advocates believe) genuine attempts to understand a phenomenon, or (as counter-advocates believe) a simple dismissal of what would otherwise be considered a failure. Wiseman provides two cases for us to examine.

His first piece of evidence for retrospective nullification of null results is a paper by Kanthamani and Broughton (KB; 1994), which reported an attempt to replicate the ganzfeld effect that yielded null results. Wiseman criticizes them for making no mention of the null hypothesis as an explanation for their nonsignificant findings, instead concluding that "it is probably safe to say that static picture targets remain a less than ideal choice for ganzfeld experiments" (Wiseman, 2010b). It is clear that Wiseman is implying that this decision was arbitrary and unwarranted, but evidence from the paper and from previous analyses contradicts his conclusion. Bem and Honorton (1994) report, for example, that among the 28 pre-PRL studies, 9 used "dynamic" targets (View Master slide reels) as compared to static pictures, and those 9 found a significantly higher hit rate than the other 19 (50% vs. 34%, respectively; Fisher's exact $p = .04$, two-tailed). Honorton's own PRL studies (Bem & Honorton, 1994) compared 164 dynamic targets to 165 static targets and also found a significant difference in scoring rates (37% vs. 27%, respectively; Fisher's exact $p < .04$). Therefore, when KB found a 27.6% hit rate in the 350 trials for their static targets, they wrote that they had replicated the finding by Bem and Honorton—and they had, to a very precise degree.

KB also found that the four groups of participants in their database that conformed to one of the four measures of "optimal subjects" as defined by Honorton (1997)—previous psi experiences, previous psi testing, a feeling-perception (FP) personality on the Myers-Briggs Type Inventory, and practice of a mental discipline—produced overall hit rates ranging from 31% to 36%. This finding is of significant importance considering that this same subpopulation aggregate for the PRL and FNRM databases—the latter an independent replication of the PRL trials (Broughton, Kanthamani, & Khilji, 1989)—was 31% (Honorton, 1997). Moreover, when three of these optimal-participant measures were combined in the KB studies, forming what Honorton (1997) called the "three-predictor model," the results were striking: KB's database exhibited a hit rate of 41.3% (46 trials; exact binomial $p = .011$, one-tailed), whereas the PRL and FNRM databases yielded a combined rate of 43% (99 trials; exact binomial $p = .0004$, one-tailed). It should be noted that these results are surprisingly consistent, and not post hoc data selection; Honorton and Schechter (1987) originally found these predictors in the PRL-1 novice series before Honorton (1997) applied them to the PRL-2 novice series, as well as the independent FNRM database, shortly before his passing. Honorton (1997) wrote:

> At the 1986 PA Convention, Honorton and Schechter (1987) presented an exploratory analysis of performance correlates for the first two PRL novice series (Series 101-102; hereafter designated PRL-1), suggesting that initial ganzfeld ESP performance was positively and significantly related to self-reports of personal psi experiences, Feeling/Perception (FP) preferences on the MBTI, and prior participation in nonganzfeld psi experiments. A positive but nonsignificant tendency for better performance among participants reporting involvement with mental disciplines such as meditation was also found. . . . In this paper, the PRL-1 findings will be compared with those in the later PRL novices series (Series 103-105; hereafter designated PRL-2) and the FRNM series to estimate the overall magnitude and consistency of the four predictors. (p. 143)

Here we should note that Honorton produced a "three-predictor model" in addition to his four-predictor model; the former was created because of the small number of subjects satisfying the prior psi testing condition, and omitted this requirement.

Recall now the results that KB found for their three-predictor dataset; if these are added to the total PRL and FRNM databases, there are 145 trials which yield a 42.06% overall hit rate (exact binomial $p = 5.07 \times 10^{-5}$, one-tailed). As for the omitted characteristic, Kanthamani and Broughton (1994) stated that prior psi testing was also successful, but because of the broader scope of the three-predictor model, they chose to apply it instead. This rather strongly confirms the improved performance of the selected participants, and it provides corroboratory evidence against the null hypothesis—even in light of the fact that the KB database *overall* is nonsignificant.

Because of these considerations, we argue that Kanthamani and Broughton (1994) were fully justified in noting that their studies confirmed the "PRL success model" (p. 7) in such a way that their conclusions cannot

be seen as evidence of retrospective nullification. Additionally, their results were not excluded from any relevant meta-analyses (Bem, Palmer, & Broughton, 2001; Milton & Wiseman, 1999; Storm et al. 2010), so even if their conclusions had been little more than confirmation bias, that would have had no effect on the evidence.

Wiseman's next example of retrospective nullification mentions Melvyn Willin's (1996a) study with musical targets as a prime model for "data mining," which Wiseman defines as the tendency to search in the results of a null study for any correlation that can yield anomalous findings. Wiseman (2010a) criticizes Willin for his decision not to invoke the null hypothesis as an explanation for his failures:

> Willin conducted a series of post hoc analyses, exploring, for example, the relationship between participants' psi scores and their age, profession, hobbies, previous paranormal experiences, and relationship with the person acting as the sender. Additional analyses explored psi scoring as a function of the month and time of day each trial was conducted. Most of these analyses yielded inconclusive results, but Willin eventually found that trials conducted early in the experiment obtained a higher hit rate than those conducted later and suggested that this might have been due to "less interest being shown by the Receivers and the Senders or by an unintentional goat effect being displayed by the Experimenter." (p. 38)

To counterbalance this information, it should be noted that Willin (1996a) began his paper with the sentence "experiments using actual music as the target have not been conducted very often" (p. 1), afterward listing several small exploratory attempts to elicit psi from musical targets, with mixed success (Altom & Braud, 1976; George, 1948; Keil, 1965; Shulman, 1938). This suggests that Willin knew his study was of a more exploratory than confirmatory nature; looking for trends and patterns after the fact was thus part of its design—especially given that it was the first large-scale experiment to employ musical targets (Parra & Villanueva, 2004). He even collected extensive background and personality data on his participants pre-analysis, for that precise purpose (Willin, 2005). We believe that post hoc findings are essential to science, so long as they are not counted as confirmatory (and the proper corrections for multiple analysis are applied), so we see no problem with his strategy. As for Willin's attitude towards his null results, we suggest that the following comment from his follow-up study (Willin, 1996b), using previously high-scoring unselected participants, should be considered: "A chance hit rate of 25% was expected and a hit rate of exactly 25% was achieved. . . . These results thus provide no evidence for the communication of music by ESP" (Willin, 1996b, p. 103).

Although we do not doubt that there are instances of confirmation bias in the parapsychology literature, wherein researchers have perhaps given undue emphasis to a success while marginalizing a failure, our review of the two situations presented by Wiseman suggests the need to be critical of such claims when they do arise, as prospective examples of confirmation bias are themselves susceptible to confirmation bias.

Expectancy Effects in Parapsychology

In addition to his critique of Kanthamani and Broughton (1994) and Willin (1996a), Wiseman (2010a) briefly mentions experimenter effects: "Perhaps the most far-reaching version of this 'get out of a null effect free' card involves an appeal to the 'experimenter effect,' wherein any negative findings are attributed to the psi-inhibitory nature of the parapsychologist running the study" (p. 37).

Since Wiseman does not give readers much information about these effects, one is left with the impression that something is fundamentally wrong with parapsychology. But Wiseman does not mention that experimenter expectancy effects have been the subject of widespread research outside parapsychology (Rosenthal, 1976). It has been known for decades that subtle psychological variables such as Rosenthal and Jacobson's (1992) "Pygmalion effect," for example, can strongly impact participant performance and affect research in nearly all behavioral fields.

Given this information, it should not be surprising that the same effects occur in parapsychology. There is a long history of studying them in the literature (Smith, 2003). Wiseman himself was party to an experiment that tested the idea and found evidence for it (Wiseman & Schlitz, 1997). The study was of the "psychic staring" effect, in which half of the trials were conducted with Wiseman (a purportedly psi-inhibitory experimenter) as the experimenter/starer, and half with Marilyn Schlitz (a purportedly psi-facilatory experimenter) in that role. Results were as predicted for the first and second collaboration: Wiseman found nothing and Schlitz found a small but statistically significant difference between the stare and no stare conditions.

Regarding this affair, journalist Guy Lyon Playfair (2014) wrote:

> In the October 2002 issue of *The Paranormal Review*, Caroline Watt asked each of them [Wiseman and Schlitz] what kind of preparations they make before starting an experiment. Their answers were: Schlitz ". . . I tell people that there is background research that's been done already that suggests this works . . . . I give them a very positive expectation of outcome. Wiseman: "In terms of preparing myself for the session, absolutely nothing."

It does not help Wiseman's case that he wrote, after the fact, that the testing process was "an enormously boring experience" (Watt, Wiseman, & Schlitz, 2002, p. 21) and that in most of the trials he was "pretty passive about it" (p. 22). However, the last collaboration failed to detect an effect with either experimenter (Schlitz, Wiseman, Watt, & Radin, 2006). At this point, both Schlitz and Wiseman reported feeling "burnt out" with the project, which, according to the experimenter expectancy effect hypothesis, *could* have led to a reduced performance for both Schlitz's and Wiseman's participants. Nevertheless, this interpretation should be taken cautiously, as it is retrospective. Of more interest is the fact that around 67% (two out of three) of Schlitz's studies achieved significant effects, while 0% (none out of three) of Wiseman's did. The contrast between their results becomes even more noticeable when the previous studies of these researchers (Wiseman et al., 1995; Wiseman & Smith, 1994; Schlitz & LaBerge, 1994) are included; Schlitz obtained a 75% (four out of five) success rate and Wiseman still had a 0% (none out of five) success rate.

Another interesting piece of evidence for a real and relevant effect comes from Smith's (2003) discussion of a study by Judith Taddonio (1976):

> Taddonio (1976) told student experimenters that the ESP test they were to use was a recently developed technique developed by Taddonio's colleagues and that the students were being asked to conduct a replication of their findings. Taddonio manipulated the expectancy of experimenters by telling those in one group that participants in previous studies using this new technique had consistently obtained above chance scores. These experimenters were assured that the test could not fail and that the results of the student's replication would give the same high scores. Experimenters in the second group were told that Taddonio's colleagues who had developed the test were worried about it because participants were all scoring well below chance. They were led to believe that the test seemed to elicit psi-missing rather than psi-hitting and that there was no doubt that the student's replication would show the same level of low scoring. In both a pilot study and a confirmatory study, participants tested by the experimenters given the positive expectancy about the test scored significantly higher than participants tested by the experimenters given the negative expectancy. (Smith, 2003, p. 75)

Taddonio's results seem to support an explanation based, again, on experimenter-participant interaction; as she wrote, a "difference in the psychological impact of the two experimenter groups upon their subjects" (Taddonio, 1976, p. 113). Confirmation bias alone seems unable to account for these experimental differences.

On a similar note, regarding the possibility that experimenter effects result from motivated scoring errors, unconscious presence of flaws, and other similar explanations, there is good news for the ganzfeld. As Robert Rosenthal wrote in 2009 (cited by Carter, 2010b):

> Ganzfeld research would do very well in head-to-head comparisons with mainstream research. The experimenter-derived artifacts described in my 1966 (enlarged edition 1976) book *Experimenter Effects in Behavioral Research* were better dealt with by ganzfeld researchers than by many researchers in more traditional domains. (p. 95)

Recent evidence supports this assertion for ESP research. Storm et al. (2010) compared the effect sizes obtained by different experimenters/laboratories in their 45-study ganzfeld and nonganzfeld noise reduction database, uncovering no significant differences between the groups using a standard ANOVA analysis ($p = .32$). For forced-choice studies, Storm et al. (2012) used the same analysis and found no evidence of a difference ($p = .36$). Mossbridge et al.'s (2012) presentiment meta-analysis, on the other hand, included no such comparison, probably

because the level of homogeneity in their database (I^2 = 27.4) obviated the necessity for one. Based on the most up-to-date meta-analytic data, then, it is by no means clear that experimenter effects present a visible problem for psi research.

In a further study of experimenter effects, Watt and Nagtegaal (2004) found that, among all the sciences examined, parapsychology had taken the strongest precautions against experimenter effects by conducting 79.1% of its research using a double-blind methodology (compared to 0.5% in the physical sciences, 2.4% in the biological sciences, 36.8% in the medical sciences, and 14.5% in the psychological sciences). These findings are consistent with those of an earlier survey on experimenter effects by Sheldrake (1998).

To conclude, although it is recognized that experimenter effects may influence parapsychological results, it should now be seen that their occurrence in the field is not unique. Neither is the presence of these effects enough to dismiss the validity of the research, for whether an experimenter is psi-inhibitory or psi-conducive, ultimately all parapsychology studies will be included in attempting to draw conclusions about psi. Most importantly, we do not believe the presence of the experimenter effect is an attempt at retrospective nullification by parapsychologists; we think the issue is a good bit more subtle than this. Such effects are, rather, inevitably involved in the process of studying and understanding a phenomenon whose properties are not fully known. Through continued research we may yet get a better grasp of them.

The Milton and Wiseman Meta-Analysis

Wiseman's next criticism invokes the Milton and Wiseman (MW; 1999) meta-analysis, which found null results for psi across all post-PRL studies conducted until 1997 and spurred a significant debate in the parapsychology community about replication (Schmeidler & Edge, 1999). Before beginning to dissect its conclusions and methodology, however, we note that, as of the most recent meta-analysis (Storm et al., 2010), the overall hit rate of the post-PRL database remains highly significant. With this in mind, it is possible to visually gauge the impact of MW's analysis by examining a plot of *z* scores across the ganzfeld, including the MW dataset (Figure 1).

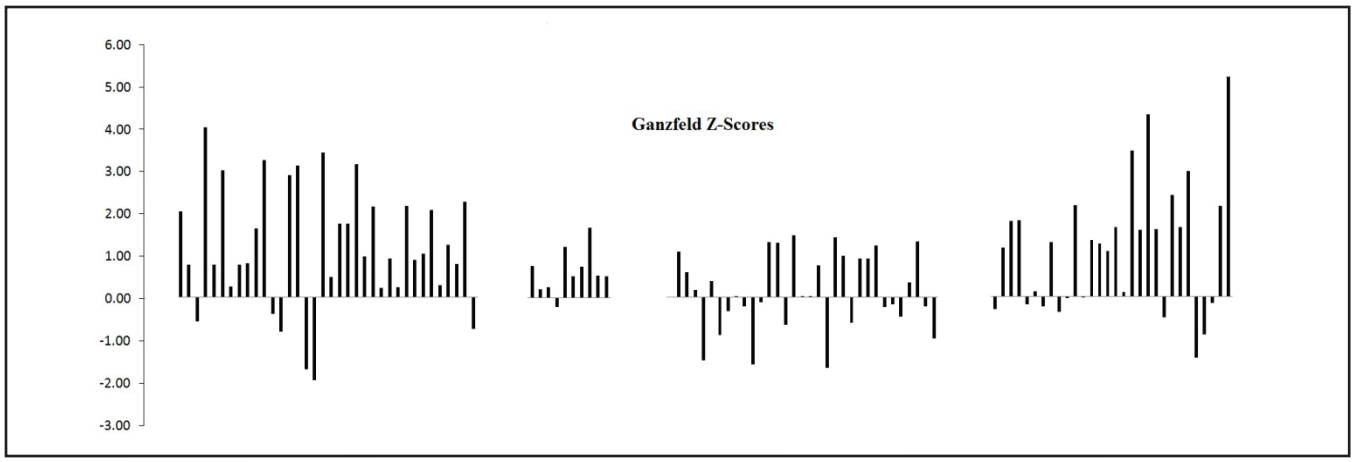

Figure1. Ganzfeld *z* scores are arranged into distinct nonoverlapping periods of research, with one exception being the PRL database, which consists only of studies by Honorton's lab. From left to right we divide the plot into pre-PRL (1974–1987), PRL (1983–1988), MW (1988–1997), and post-MW (1997–2008). Eleven overlooked studies have been added to the pre-PRL and MW databases, found for those periods by Storm and Ertel (2001).

As we can see in Figure 1, the period from 1988–1997, during which the MW meta-analysis was conducted, was the most troubling for the ganzfeld. We will explore reasons for this as well as explain why the MW database is not as negative as it initially seems.

First, if one performs an exact binomial test, the most accurate test of statistical significance, the overall result of the MW meta-analysis is marginally significant (p = .026, one-tailed). Milton and Wiseman used the less accurate standard unweighted Stouffer *Z* method, which gives the *z* score of a four-trial study as much influence on the overall *z* as the *z* score of a 128-trial study. It should be noted that some other meta-analyses used this same

analysis (e.g., Bem et al., 2001; Honorton, 1985; Storm et al., 2010), although Storm et al. did present the binomial test as a supplementary analysis for their 30-study database from 1997–2008, and found highly significant overall outcomes for the 29 studies using a four-choice design. Nevertheless, the Bem and Honorton (1994) meta-analysis, which Milton and Wiseman were attempting to replicate, obtained its highly significant overall hit rate using the binomial test. Whatever one's conclusions about the validity of MW's statistical approach, we would argue that any meta-analysis claiming as its goal the attempted replication of an earlier meta-analysis on the same type of studies should use the same statistical test of significance, especially when that test of significance is the most accurate one available.

We acknowledge, however, that choice of statistical test does not change the precipitous drop in ES observed during the period that MW's analysis covered. One possible explanation for this is offered by Bem et al. (BPB; 2001):

> The z scores of the studies in the Milton-Wiseman database are significantly heterogeneous, and one of the observations made during the online debate was that several studies contributing negative *z* scores to the analysis had used procedures that deviated markedly from the standard ganzfeld protocol. Such a development is neither bad nor unexpected. Many psi researchers believe that the reliability of the basic procedure is sufficiently well established to warrant using it as a tool for the further exploration of psi. Thus, rather than continuing to conduct exact replications, they have been modifying the procedure and extending it into unknown territory. Not unexpectedly, such deviations from exact replication are at increased risk for failure. For example, rather than using visual stimuli, Willin modified the ganzfeld procedure to test whether senders could communicate musical targets to receivers. They could not. When such studies are thrown into an undifferentiated meta-analysis, the overall effect size is thereby reduced, and perversely, the ganzfeld procedure becomes a victim of its own success. (p. 208)

Rather than attempt to resolve what constituted "standard" ganzfeld procedure, BPB took the experimental route. They tasked three blind raters, each unfamiliar with the study outcomes, to rate the 40 studies in their database (from which all results had been erased), according to a 7-point standardness scale (where 7 indicated the greatest conformity to PRL protocols, and 1 indicated the least). As their guide to defining standardness, they were given the two original PRL ganzfeld papers of Bem and Honorton (1994) and Honorton et al. (1990).

However, Wiseman writes that BPB added a standardness measure that was not found in the papers given to the blind raters: participant selection. Measures such as "prior meditation experience," "artistic or creative," and "mental discipline practice," Wiseman claims, were post hoc conditions based on knowledge of experimental outcomes, and therefore examples of retrospective data selection. But Wiseman is likely misled on this point. Bem and Honorton (1994) make it clear that replications should use participants of the selected type, probably because 100% of the PRL participants were selected in at least one of the previously mentioned ways, or others (e.g., strong belief in psi, friends, biologically related). Wiseman additionally went on to say, in a 2010 talk based on his article (Wiseman, 2010b), that the use of no sender was also treated as standard, even though the ganzfeld was considered a telepathy methodology. However, he did not mention that the PRL studies gave the participants the option of whether or not to have a sender, and four participants opted to have no sender. Honorton never claimed the ganzfeld method was only for testing telepathy; he would tell participants that it was for testing telepathy because it would seem more plausible to them than clairvoyance or precognition, therefore giving them more motivation. Donald McCarthy (1993) wrote of Honorton:

> He told me, not long ago, that in designing the ganzfeld procedure, a primary reason for his choosing a telepathy protocol was that it might lead to more ready acceptance, since people seemed less threatened by the idea of "mental radio" than by other ways of conceptualizing psi. (p. 9)

The end result of the BPB analysis was strikingly in accordance with expectation; the success of the studies correlated significantly with the measures used to evaluate compliance with PRL protocols. Those studies that ranked above 4.0 on the scale (the midpoint) yielded significant results at a hit rate of 31.2% (1,278 trials, 29 studies, exact binomial $p = .0002$, one- tailed), and those that fell below gave a hit rate of 24% (n.s.). More dramatically, the studies that went to 6 or above (974 trials, 21 studies) scored a 33% hit rate (exact binomial $p = 1.58 \times 10^{-8}$, or odds

against chance of 63 million to 1)—almost exactly PRL's own. We argue that this is another example of a surprising consistency in psi research.

A possibility, however, that may have confounded the conclusions of the BPB meta-analysis, pertains to selected participants. Because Milton and Wiseman never did a heterogeneity analysis, they did not notice that their database was significantly heterogeneous by Timm's chi square test on *z* scores, and significantly heterogeneous by Honorton's chi square test on effect sizes ($p = .07$, two-tailed; the alpha for chi square tests on meta-analyses is $p \leq .05$ to compensate for low statistical power). This led one of us (Derakhshani) to test the hypothesis that the source of the heterogeneity might have been the difference in scoring rates between selected and unselected participants. Given that the PRL database used only selected participants, it could be argued that a meaningful indicator of replication would be a comparison between the hit rates of the PRL studies and the MW studies that used selected participants. In fact, for the 513 trials in MW's database from studies that used selected participants, 157 hits were obtained, for an overall hit rate of 30.6% (exact binomial probability of $p = .002$, one tailed), which is not significantly different from the 32.2% hit rate of the PRL studies (Fisher's exact $p = .65$, two tailed). By contrast, the 661 trials with unselected participants produced a 24.7% overall hit rate, which is significantly different from that of the selected participants (Fisher's exact $p = .014$, one-tailed) and, incidentally, also nonsignificantly different from the 27.3% hit rate of unselected participants in Storm et al. (2010) with four-choice designs; $p = 0.40$. Thus, it can be argued that Milton and Wiseman (1999) actually replicated the PRL results, if one considers roughly homogenous populations alone—without even factoring in the relevance of homogeneous procedures—although the greater accuracy obtained through the standardness ratings in BPB suggests that study procedures still play a role in moderating outcomes.

In any case, as Storm et al. (2010) and many other meta-analyses (Bem et al., 2001; Derakhshani, 2014; Radin, 2006; Storm & Ertel, 2001; Storm et al., 2010; Tressoldi, 2011; Utts, Norris, Suess, & Johnson, 2010) demonstrate, the overall hit rate of the post-PRL database remains highly significant after the MW meta-analysis. Dean Radin (2010) makes an important point about this—namely that the controversy over replication in the ganzfeld (and in other psi paradigms) has advanced beyond the replicability of individual studies and is now about the replicability of experiments considered in groups. Not only are there individual meta-analyses confirming (what we have argued to be) a reasonable rate of replication across a wide swath of experiments over periods of several years; there are now groups of meta-analyses confirming consistency over many thousands of trials, in more than a hundred studies, and over five decades.

The Declining Decline Effect

In the last section of his article, Wiseman (2010a) writes:

> The alleged psi effects associated with a certain procedure frequently have a curious habit of fading over the course of repeated experimentation. Skeptics argue that this is due to the parapsychologists identifying and minimizing potential methodological and statistical flaws over time. However, some parapsychologists have come up with creative ways of explaining away this potential threat, arguing that such decline effects are either an inherent property of psi or that psychic ability really does exist but is inversely related to the level of experimental controls employed in a study. (pp. 37–38)

As in the example of experimenter expectancy, we believe that Wiseman has left out several important observations for the decline effect. Moreover, we have already pointed out that the hypothesis from the preceding quote that "psychic ability . . . is inversely related to the level of experimental controls employed in a study" is not supported by the evidence available on a meta-analytic level, for most of the kinds of effects examined by experimental psi research. Additionally, in tandem with experimenter expectancy, decline effects are far from unique to parapsychology.

For example, Jonathan Schooler (2011), professor of psychological and brain sciences at the University of Santa Barbara, covered a number of examples of the decline effect in a debate over psi at Harvard, showing that they occur in research on schizophrenia, with medicines such as Pravastatin, Timolol, and Latonoprost, and even ecological relationships. Journalist Jonah Lehrer (2010) also wrote about the decline effect in a controversial article published in the *New Yorker*, which discussed the phenomenon's occurrences in tests of the drug Zyprexa,

psychological effects such as verbal overshadowing (Schooler's own widely cited research), biological correlations between asymmetry and mutation, and other paradigms. In order to contextualize the decline effect, Schooler (2011) proposed the following framework during a debate about psi at Harvard:

Controversial Prediction

Conceptual extensions in new domains will initially work but will then similarly decline

Mainstream accounts

Regression to the mean

Refinement of procedure

Confirmatory rather than exploratory research

Controversial account

Heisenberg effects generalize in some yet unknown manner to scientific observation of phenomena

Genuine effects actually fade with repeated observation

** Mainstream view clearly most parsimonious at present but uncertain until decline effect is adequately understood. Need a process for recording all negative and unpublished findings to resolve issue*

An additional hypothesis for the decline effect not touched upon by Schooler in the above framework is the possibility of increased publication bias around the time an effect is first reported to be produced—as mentioned by Lehrer (2010). Indeed, Harris and Rosenthal (1988) illustrate this: They predicted in their assessment of Honorton's (1985) ganzfeld meta-analysis that, taking into account corrections for minute publication bias, along with corrections for statistical errors and reporting errors, the true ganzfeld hit rate would decrease from 38% to around one third. This prediction was strikingly confirmed by the PRL meta-analysis (Bem & Honorton, 1994), which found a 32% overall hit rate across its 10 studies.

So the skeptical hypothesis mentioned by Wiseman for the decline across time is likely close to the truth for the drop in effect size from the earliest ganzfeld database to the second. But does it account for the decline in effect size from the PRL to the Milton and Wiseman (1999) database? Although declines can come about for a number of reasons, as regards the MW database two possible explanations are (a) treatment of exploratory research as confirmatory and (b) a change in the population tested over time. There is strongly supportive evidence for both. However, there is no available evidence that this decline can be explained by higher-quality research in the MW database, relative to the PRL database. This explanation also fails to account for the subsequent incline from the MW database to the post-MW database, or the significant incline across both, if they are considered together: $r = .27$, $p = .03$.

What about just the ganzfeld studies after the MW meta-analysis? To address this, we looked at data from the most recent post-MW ganzfeld database of 30 studies, from 1997–2008. We examined study ES versus study year and study ES versus study quality ratings, both for the entire database of 30 studies and for the two homogeneous subgroups of selected and unselected participants identified by Derakhshani (2014).

In plotting study effect size vs. study year, we found no decline in effect size ($r = 0$). Plotting study ES vs. study quality ratings, we found a *positive and significant* correlation, $r(28) = .37$, $p = .045$, two-tailed. That is, studies rated as having higher methodological quality produced *larger* effect sizes than lower quality studies, and this trend was statistically significant. For the entire database, then, the evidence appears to contradict Wiseman's hypothesis.

Of course, one might reasonably argue that if the entire database is heterogeneous in ES distribution, then the correlations involving effect size may be misleading. Derakhshani (2014) indeed found highly significant

heterogeneity (p = .002, two-tailed) via the chi square test. He also found that blocking the studies according to whether they used selected or unselected participants produced two safely homogeneous databases. The selected participant subgroup consisted of 14 studies of four-choice design, with a 40.1% overall hit rate across 748 trials; and the unselected participant subgroup consisted of 15 studies of four-choice design (Roe & Flint, 2007, was excluded because it used an eight-choice design) with a 27.3% overall hit rate across 886 trials. Furthermore, the difference in hit rates was extremely significant, Fisher's exact $p < .0001$).

A reasonable explanation for this is that selected studies have lower average quality ratings, but as Derakhshani (2014) shows, this is not the case. Selected participants in fact produced a sample size weighted mean quality rating of q = .84 (where 1 is the highest possible rating), whereas the unselected participant studies produced a lower mean quality rating (q = .79).

We did find that there was a small negative correlation between ES and study year for the selected participant studies, but it was not significant, $r(12) = .30$, $p = .29$, two-tailed. Moreover, we found a positive, nonsignificant correlation, $r(12) = .27$, $p = .37$, two-tailed, between study quality and study ES. We also found a positive and nonsignificant correlation, $r(12) = .26$, $p = .37$, two-tailed, between study quality rating and study year. Our analyses thus do not support a relationship between quality and ES. More selected studies will be needed before we can ascertain whether the positive correlation between these variables is real or spurious.

The unselected participant subgroup, on the other hand, had more striking results. For ES vs. year, we found a *highly significant* positive correlation, $r(14) = .65$, $p = .007$, two-tailed. For quality ratings vs. year, we found an *extremely significant* positive correlation, $r(14) = .86$, $p < .00002$, two-tailed. And for quality ratings vs. ES, we found a large but nonsignificant positive correlation, $r(14) = .40$, $p = .13$, two-tailed.

On the basis of these results, we can say with confidence that there is no decline in ES for unselected studies, and that there is no evidence for one in selected participant studies. We admit that we do not know why the findings are so robust for the unselected participant subgroup, and nonsignificant for the selected participants subgroup, but the question surely merits further research.

Given our reliance on the quality criteria of Storm et al. (2010), however, a skeptical reader might reasonably ask if there could be something problematic with or implausible about how these were constructed/judged. Storm et al.'s (2010) quality ratings were made by two judges (graduate students of Tressoldi) who saw only the method section of each article they assessed; all identifiers had been deleted, such as article title, authors' hypotheses, and references to results of other experiments. The seven criteria by which they evaluated the quality of a study are reproduced for convenience below (Storm et al., 2010, p. 474):

1. Appropriate randomization (using electronic apparatuses or random tables).
2. Random target positioning during judgment (i.e., target was randomly placed in the presentation with decoys).
3. Blind response transcription or impossibility to know 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 blind to target identity.

Two judges answered "yes" or "no" to each of the criteria. Study quality was 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), and the quality ratings of each judge were averaged together. Storm et al. (2010) reported a Cronbach alpha for the two judges' 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 can see no major flaws in them that might nullify our conclusions, for either the unselected or selected participants subgroup.

To summarize, the selected participants subgroup shows no evidence for the skeptical hypothesis, and the unselected participant subgroup shows some evidence against. As we can find nothing evidently wrong with the criteria and methods by which Storm et al. (2010) determined study quality, we conclude that the skeptical hypothesis suggested by Wiseman is inconsistent with the data, at least when it comes to the ganzfeld paradigm after the PRL experiments.

Beyond the ganzfeld, the forced-choice ESP meta-analysis by Storm et al. (2012) found a *positive and highly significant* incline effect for study year vs. study ES in their homogeneous database of 72 studies from 1987–2010, $r = .31$, $p = .007$, two-tailed, along with a *positive and significant* correlation between study year and study quality rating in their heterogeneous 91-study database, $r = .25$, $p = .02$, two-tailed. They also found a *very weak, negative, and nonsignificant* correlation between quality rating and ES, $r = -0.08$, $p = .45$, two-tailed. Honorton and Ferrari's (1989) assessment of the forced-choice precognition literature, likewise, found that ESs had remained relatively constant through 1936–1987, although quality had substantially improved.

In sum, we find little evidence in either ganzfeld or forced-choice experiments for problematic decline effects. There is significant evidence in recent ganzfeld work for an incline across the MW database to the post-MW database, exactly no incline among all post-MW studies, and highly significant evidence for an incline just within unselected subjects studies for the post-MW database. As we have noted, interesting questions remain to be pursued, such as (a) what are the true correlations for ES vs. quality ratings, ES vs. year, and quality ratings vs. year, in the ganzfeld selected participant subgroup, and (b) why are the correlations so strongly significant for the unselected participants subgroup but nonsignificant for the selected participants subgroup? Such research questions, however, were not present in Wiseman's discussion of declines.

The Progress of Parapsychology

Below, we present Wiseman's perspective on the history of psi research and contrast it with our own. Wiseman (2010a) states:

> Initial work, conducted between the early 1930s and late 1950s, primarily involved card guessing experiments in which people were asked to guess the identity of specially printed playing cards carrying one of five simple symbols. By the mid-1960s parapsychologists had realized that such studies were problematic to replicate and so turned their attention to dream telepathy and the possibility of participants predicting the outcome of targets selected by machines. In the mid 1970s and early 1980s, the ganzfeld experiments and remote viewing took over as dominant paradigms. In 1987, a major review of the area by parapsychologists K. Ramakrishna Rao and John Palmer argued that two sets of ESP studies provided the best evidence for the replicability of psi: the ganzfeld experiments and the differential ESP effect (wherein participants apparently score above chance in one condition of an experiment and below chance in another). More recently, parapsychologists have shifted their attention to alleged presentiment effects, wherein participants appear to be responding to stimuli before they are presented. Finally, there are now signs that the next new procedure is likely to adopt a neuropsychological perspective, focusing on EEG measurements or functional MRI scans as people complete psi tasks. (p. 39)

According to our assessment of the literature, the reason for the shift in research focus from (for example) dream ESP to the ganzfeld, and from the ganzfeld to presentiment, is the goal of finding an experimental paradigm that produces the largest ES for the least financial cost and time per trial. The ganzfeld, for example, is well known to produce comparable ESs to dream ESP but for a fraction of the time and cost per session (a typical ganzfeld trial typically takes 1–2 hours, compared to a full 24 hours for a dream ESP trial), and presentiment produces comparably greater effect sizes to ganzfeld *on average*; random effects ES = 0.13 (Tressoldi, 2011) and random effects ES = 0.21 (Mossbridge et al., 2012), respectively—but for even less time and cost per trial (typically a few seconds or minutes compared to 1–2 hours).

Even so, we emphasize that each experimental approach has its advantages—both for producing psi and understanding it—and that research in other paradigms certainly has not ended. The recent meta-analysis of forced-choice ESP studies by Storm et al. (2012) shows, for example, that 91 studies of admissible methodological quality were conducted from 1987–2010. By comparison, for the ganzfeld, only 60 such studies were found in the Storm et

al. (2010) meta-analysis, from 1987–2008. Despite the rise of the presentiment paradigm in the late 1990's, moreover, Storm et al. still reported 30 ganzfeld studies conducted in the period from 1997–2008, exactly matching the number of studies in the Milton-Wiseman (1999) database from before the rise (1987–1997). These observations suggest that some research paradigms have been only minimally affected by the emergence of others.

In sum, our analysis leads us to reject both of Wiseman's claims about psi research: that (a) decline effects consistently lead to nonreproducible results, and (b) parapsychologists routinely abandon old experimental procedures for new ones. On the contrary, we found evidence that research continues in the majority of parapsychology paradigms, and we think that there are more persuasive reasons than replication failure for why parapsychologists work to pioneer (and adopt) new research techniques.

A Couple Suggestions for the Way Forward

In Wiseman's (2010a) concluding section, he writes:

> To help the field move forward and rapidly reach closure on the psi question, parapsychologists need to make four important changes in the way they view null findings. First, they should stop trying lots of new procedures and cherry-picking those that seem to work and instead identify one or two that have already yielded the most promising results. Second, rather than varying procedures that appear successful, they should instead have a series of labs carry out strict replications that are both methodologically sound and incorporate the most psi-conducive conditions possible. Third, researchers should avoid the temptation for retrospective meta-analysis by pre-registering the key details involved in each of the studies. And finally, researchers need to stop jumping ship from one experimental procedure to another and instead have the courage to accept the null hypothesis if the selected front-runners don't produce evidence of a significant and replicable effect. (p. 39)

Although we hope to have shown that the charges of "cherry-picking" new procedures—on the basis of the examples examined—are questionable, we do agree with Wiseman that parapsychology could benefit from focusing its resources on fewer research paradigms and using the best meta-analytic data on these paradigms to boost ESs and replication rates as high as possible. Here, we would like to make our own humble suggestions for how this could be done, according to our review of the evidence.

First, Derakhshani (2014) advises, on the basis of a predictive power model utilizing existing meta-analyses of ganzfeld studies, that it should be possible to boost the replication rates of future ganzfeld studies from ~30% to as high as 80%, while keeping the mean sample sizes of ganzfeld studies effectively the same, by the careful and exclusive use of selected participants in all (or as many possible) future ganzfeld studies. We want to be explicit: For the most recent Storm et al. (2010) database, the selected participant hit rate of 40% suggests that the recipe exists for a ganzfeld study with a significantly amplified effect size and chance of success. But care is necessary. This hit rate differs greatly from that of selected participants from previous databases (30% for the MW and 32% for the PRL) and therefore is likely not explicable by just any selection process.

Although we have not conducted an exhaustive review of the differences between the present selected population and previous selected populations, the success of participants in the PRL, FNRM, and KB databases who satisfied Honorton's (1992) three-predictor model—previous psi experience, a feeling-perception (FP) typology on the Myers-Briggs Personality Inventory, and practice of a mental discipline—exceeded the success of participants satisfying only one of the four optimal participant traits identified by Honorton and Schechter (1987), by 42% to 31%. This result suggests *strongly* that combinations of such traits are superior to just one or two. Indeed, given the superior performance of the three-predictor model, we suggest that it would be a reasonable time to retest it. For participants satisfying the three-predictor model—if we assume the hit rate across the PRL, FNRM, and KB databases (42% across 143 trials)—the sample size required for 80% power is only 48 trials.

In a similar vein, using Storm et al.'s (2010) selected participant hit rate of 40%, the required sample size is just 56. We emphasize that the characteristics of this selected sample have not been systematically reviewed; nevertheless, we can identify two highly powered studies in Storm et al. (2010) that provide a model for future investigators. Dalton (1997), using preselected artistic participants with positive attitudes towards psi and previous psi experiences, obtained a 47% hit rate in 128 trials (and also had the highest quality rating of 1.00 in Storm et al's

2010 meta-analysis). Parra and Villanueva (2006), who used participants that were mostly psi believers and reported having previous psi experiences and training in meditation, found a 41% hit rate in 138 trials. Future ganzfeld researchers would do well to emulate these studies. As a final piece of advice for ganzfeld studies, it should be noted that Derakhshani (2014) calculated the required sample size for artistic participants (across the ganzfeld databases) to reach 80% power, and it is approximately 47 trials (for a 41% hit rate in 367 trials). Artistic populations thus seem to constitute the optimum ganzfeld replication pool.

If parapsychologists keep their studies relatively standard methodologically (given the positive correlation between standardness ratings and effect sizes in Bem et al., 2001), of high quality (given the positive correlation between quality and effect size found by Derakhshani), and use only well-selected participants, we predict the possibility of replication rates of 80% or greater. Large sample sizes also seem advisable given the positive correlation between *N* and ES found by Derakhshani for selected participants studies. In our view, such success would go a long way towards persuading the mainstream to replicate these studies.

Second, with regard to nonganzfeld experiments, in the Honorton and Ferrari (1989) forced-choice precognition meta-analysis, a subset of "optimal studies" in the homogenous database using selected subjects and trial-by-trial feedback had remarkably high replication rates—of the eight optimal studies, seven (87.5%) produced significant outcomes at the .05 level, with mean $z = 6.14$ and mean ES = 0.06. In the heterogeneous version of their database, there were 17 optimal studies, 15 (88%) of which reached statistical significance at the .05 level, with combined $z = 15.84$ and mean ES = 0.12—about twice the ES of ganzfeld studies using unselected participants. By comparison, the nine sub-optimal studies in their homogeneous database produced no significant results; mean $z = -1.29$ and mean ES = 0.005; and the optimal studies had significantly higher quality ratings than the sub-optimal studies (optimal mean = 6.63, $SD = 0.92$; suboptimal mean = 3.44, $SD = 0.53$; $t(10) = 8.63$, $p = 3.3 \times 10^{-6}$, two-tailed). Furthermore, for the more recent Storm et al. (2012) forced-choice meta-analysis, one of us (Derakhshani) found 11 studies using selected participants that produced a mean ES = 0.09, with 8/11 (73%) significant at the .05 level. By comparison, the 80 studies using unselected participants produced only 21/80 (26.3%) significant studies, with mean ES = 0.03. Derakhshani also found that the mean quality rating of the studies using selected participants was lower than the mean quality rating for studies using unselected participants ($q = 0.69$ vs. $q = 0.81$), but the difference between ratings was not significant (selected participant studies mean rating = 0.69, $SD = 0.23$; unselected participant studies rating = 0.80, $SD = 0.21$; $t(89) = 1.61$, $p = .11$, two-tailed. In addition, 6 of the 11 selected participants studies produced a mean quality rating of 0.86, with no study rated less than 0.80, and yet the mean ES of 0.08 for these six studies was still more than triple the mean effect size of the unselected participants studies. In addition, Derakhshani noticed that the 11 selected participants studies produced a strong correlation between *N* and z; $r = .64$, $p = .012$, one-tailed, as would be expected under the assumptions of power analysis. By comparison, the 80 unselected participant studies produced a null correlation ($r = 0$). On the basis of these findings, we suggest that now is a suitable time for parapsychologists to reinvest in the forced-choice paradigm, prospectively plan a large set of optimally designed forced-choice ESP studies, and set the sample sizes for these studies large enough that they can expect at least 80% to reach statistical significance, assuming one of the observed effect sizes for optimal studies (we recommend using the conservative lower effect size estimate of 0.055). The expected outcome would then be that at least 80% of such studies should reach significance at the .05 level.

We also suggest that experimenters make every use of venues such as the Koestler Parapsychology Unit Registry (http://www.koestler-parapsychology.psy.ed.ac.uk/TrialRegistry.html) to preregister experiments, in order to circumvent what little publication bias may still exist in parapsychology. Additionally, we recommend careful examination of all methodological and statistical guidelines suggested by the program chairs of the Parapsychological Association 56th annual convention, originally put forward by Utts and Tressoldi (2013). Finally, we strongly recommend that any future large prospective studies make use of the two safeguards against experimenter misconduct proposed by Kennedy (2014): registering a multiple-experimenter protocol with independent copies of study outcomes, so as to prevent tampering; and providing the raw data for analysis by others after a study is completed.

If parapsychologists were to adopt these suggestions for ganzfeld and forced-choice ESP studies and produce the outcomes predicted on the basis of these meta-analytic findings, we believe it would go a long way towards convincing the mainstream academic community to take seriously the scientific possibility of ESP, as well as to invest resources to attempt large-scale replications of the results. Conversely, if the predicted outcomes were grossly disconfirmed, it would raise serious doubts about the positive results of past experiments. Either outcome, in our view, would constitute significant progress in the scientific assessment of whether or not ESP exists.

Final Thoughts

Replying to a series of critiques of psi research similar to Wiseman's, Honorton (1993) wrote that counter-advocates of parapsychology had drifted away from making active contributions to the field—some of which in the past (e.g. Hyman & Honorton, 1986) were substantial:

> Critics have been forced to admit that parapsychology has demonstrated anomalous effects that need to be explained and they have run out of plausible conventional explanations . . . instead, they offer a caricature of the history of parapsychology and present polemical arguments designed to convince us that there is really nothing in parapsychology that warrants scientific interest, except, perhaps, for the motivations of those who persist in studying it. (Honorton, 1993, p.191)

In reviewing Wiseman's essay, *Heads I Win, Tails You Lose: How Parapsychologists Nullify Null Results*, we have come to the opinion that it too presents such a caricature. Wiseman's portrait of parapsychology is simply more dismal than the data. The impression one is left with after reading it, not uncommon in published criticisms of psi research, is that parapsychology is an unprofitable area of inquiry, from which little is to be gained but frustration with the caprice of psi and the confirmation bias of psi researchers. This vignette is regrettable in our opinion, and poorly serves both parapsychology and organized skepticism; for whereas the former is depleted of its financial and human support as a result of negative publicity, the latter is—just as importantly—deprived of a unique opportunity to examine the evidence for psi, for the same reason.

We hope to have shown that the opportunity is still there. The evidence for some forms of psi is stronger than we would expect it to be; it is not easily dismissed, easily ignored, or, indeed, easily summarized. Our approach considered and responded to each general criticism raised by Wiseman with specific evidence from the literature, and showed that parapsychology has reason to intrigue all participants in the psi debate, from advocates to counteradvocates to (like us) newcomers becoming familiar with the field. On the other hand, by putting forth recommendations, we have tried to demonstrate that parapsychology can still improve its face validity. High-powered prospective studies, sizeable proportions of significant results, large effect sizes, open source dissemination of data, pre-registered studies, multiple-experimenter protocols, and much more, are perhaps just around the corner.

It is this commitment to improve, in fact, where both sides can meet, for we believe the joint goal *must be* to produce the highest quality parapsychology database within our present means—a database that we have argued is strongly suggested by the research. If we maintain a willingness to improve the evidence to this degree then perhaps the debate will advance "Beyond the Coin Toss"—dispensing with heads, tails, and the single outcomes of winning and losing, implied in the titles of Wiseman's (2010a) essay and Carter's (2010a) response, for a collaborative attempt to resolve the psi enigma for the twenty-first century.

References

Ahmed, I., Sutton, A., & Riley, R. (2012). Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: A database survey. *British Medical Journal*, *344*,d7762. Retrieved from http://www.bmj.com/search/advanced?text_abstract_title=publication%20bias&limit_from=2012-01-01&limit_to=2012-12-31&sort=relevance-rank&resultformat=standard&numresults=10&flag=research&searched=yes. doi: 10.1136/bmj.d7762

Altom, K., & Braud, W. G. (1976). Clairvoyant and telepathic impressions of musical targets [abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology 1975* (pp. 171–174). Metuchen, NJ: Scarecrow Press.

Bakker, M., Dijk, A. V., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554. doi: 10.1177/1745691612459060

Begley C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*, 531–533. doi:10.1038/483531a

Bem, D. J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, *115*, 4–18. doi: 10.1037//0033-2909.115.1.4

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.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A)*, *23*, 399–406. doi: 10.1076/jcen.23.3.399.1181

Blackmore, S. (1980). The extent of selective reporting in ESP ganzfeld studies. *European Journal of Parapsychology*, *3*, 213–219.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators. A meta-analysis. *Psychological Bulletin*, *132*, 497–523.

Brady, C., & Morris, R. (1997). Attention focusing facilitated through remote mental interaction: A replication and exploration of parameters. *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention*, 73–91.

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.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. doi: 10.1038/nrn3475

Carter, C. (2010a). 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,* 156–167.

Carter, C. (2010b). Persistent denial: A century of denying the evidence. In S. Krippner & H. L. Friedman (Eds.), *Debating psychic experience: Human potential or human illusion?* (pp. 77–110). Santa Barbara, CA: ABC-CLIO.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. doi: 10.1037/h0045186

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

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

Delanoy, D., & Sah, S. (1994). Cognitive and physiological psi responses to remote positive and neutral emotional states. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention*, 128–138.

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

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One*. Retrieved from http://www.plosone.org/article/info%253Adoi%252F10.1371%252Fjournal.pone.0010068. doi: 10.1371/journal.pone.0010068

George. R. (1948). An ESP experiment with music. *Parapsychology Bulletin*, *11*, 2–3.

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*, 363–386. doi: 10.1037//0033-2909.97.3.363

Harris, M., & Rosenthal, R. (1988). *Interpersonal expectancy effects and human performance research*. Retrieved from http://www.nap.edu/openbook.php?record_id=779&page=1

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fncom.2012.00008/abstract. doi: 10.3389/fncom.2012.00008

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

Honorton, C. H. (1993). Rhetoric over substance: The impoverished state of skepticism. *Journal of Parapsychology, 57*, 191–214.

Honorton, C. H. (1997). The ganzfeld novice: Four predictors of initial ESP performance. *Journal of Parapsychology, 61,* 143–158.

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

Honorton, C. H., & Schechter, E. (1987). Ganzfeld target retrieval with an automated testing system: A model for initial success [abstract]. In D. H. Weiner & R. D. Nelson (Eds.), *Research in parapsychology 1986* (pp. 36–39). Metuchen, NJ: Scarecrow Press.

Hyman, R., & Honorton, C. H. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology*, *50*, 351–164.

Ioannidis J. P. (2005). Why most published research findings are false. *PLoS Medicine2*, e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253. doi: 10.1177/1740774507079441

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

Keil, H. H. J. (1965). A GESP test with favourite musical targets. *Journal ofParapsychology*, *29*, 35–44.

Kennedy, J. E. (2014). *Experimenter misconduct in parapsychology: Analysis manipulation and fraud*. Unpublished manuscript.

Retrieved from http://jeksite.org/psi/misconduct.htm
Kosciulek, F., & Szymanski M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin*, *36*, 212–219.
Krippner, S., & Friedman, H. L. (2010). *Debating psychic experience: Human potential or human illusion?* Santa Barbara, CA: ABC-CLIO.
Lehrer, J. (2010). *The decline effect and the scientific method*. Retrieved from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer
McCarthy, D. (1993). To boldly go: An appreciation of Charles Honorton. *Journal of Parapsychology*, *57*, 7–23.
Milton, J., & Wiseman, R. (1999). Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin*, *125*, 387–391. doi: 10.1037//0033-2909.125.4.387
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., Summers, J., & Yim, S. (2003). Evidence of anomalous information transfer with a creative population. *Proceedings of Presented Papers: The Parapsychological Association 46th Annual Convention*, 116–131.
Mossbridge, J., Tressoldi, P., & Utts, J. (2012). Predictive physiological anticipation preceding seemingly unpredictable stimuli: A meta-analysis. *Frontiers in Perception Science*. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fpsyg.2012.00390/full. doi: 10.3389/fpsyg.2012.00390
Nosek, B. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. doi: 10.1177/1745691612462588
Nosek, B. A., Lai, C. K., LeBel, E. P., Gilbert, E., & Strohminger, N. (2014, May 23). *The reproducibility project: Estimating the reproducibility of psychological science*. Paper presented at the meeting of the Association for Psychological Science, San Francisco, CA.
Parker, A. (2000). A review of the ganzfeld work at Gothenburg University. *Journal of the Society for Psychical Research*, *64*, 1–15.
Parra, A., & Villanueva, J. (2004). Are musical themes better than visual images as ESP-targets? An experimental study using the ganzfeld technique. *Australian Journal of Parapsychology*, *4*, 114–127.
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.
Playfair, G. L. (2014). Skeptics have us covered. *Skeptical Investigations*. Retrieved from www.skepticalinvestigations.org/Observeskeptics/Observer/observer6.html
Radin, D. (2006). *Entangled minds: Extrasensory experiences in a quantum reality*. New York: Paraview Pocket Books.
Radin, D. (2010). The critic's lament: When the impossible becomes possible. In S. Krippner & H. L. Friedman (Eds.), *Debating psychic experience: Human potential or human illusion?* (pp. 113–128). Santa Barbara, CA: ABC-CLIO.
Radin, D., &Nelson, R. (2003). Research on mind-matter interactions (MMI): Individual intention. In W. B. Jonas & C. Crawford (Eds.), *Healing, intention and energy medicine: Science, research methods and clinical implications* (pp. 39–48). London: Churchill Livingstone.
Radin, D., Nelson, R., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin*, *132*, 529–532. doi: 10.1037/0033-2909.132.4.529
Rhine, J. B., Pratt, J. G., Stuart, C. E., Smith, B. M., & Greenwood, J. A. (1967). *Extrasensory perception after sixty years*. Boston: Bruce Humphries.
Richard, F. D., Bond, C. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. doi: 10.1037/1089-2680.7.4.331
Roe, C. A., & Flint, S. (2007). A remote viewing pilot study using a ganzfeld induction procedure. *Proceedings of Presented Papers: The Parapsychological Association 50th Annual Convention,* 198–201.
Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Wiley.
Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom*. New York: Irvington.
Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 Years? *Journal of Consulting and Clinical Psychology, 58*, 646–656. doi: 10.1037/0022-006X.58.5.646
Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: Wiley. doi: 10.1002/0470870168.ch1
Scargle, J. D. (2000). The file-drawer problem in scientific inference. *Journal of Scientific Exploration*, *14*, 91–106.
Schlitz, M., Wiseman, R., Watt, C., & Radin, D. (2006). Of two minds: Sceptic-proponent collaboration in parapsychology. *British Journal of Psychology, 97*, 313–322. doi: 10.1348/000712605X80704
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., Schneider, R., Utts, J., & Walach, H. (2004). Distant intentionality and the feeling of being stared at: Two meta-analyses. *British Journal of Psychology*, *95*, 235–247. doi: 10.1348/000712604773952449

Schooler, J. (2011). *Reflections on the pursuit of psi*. Lecture presented at Harvard University, Cambridge, MA. Retrieved from http://www.wjh.harvard.edu/~moulton/psi_panel_schooler.ppt

Schouten, S. (1993). Are we making progress? In L. Coly & J. McMahon (Eds.), *Psi research methodology: A re-examination* (pp. 295–332). New York: Parapsychology Foundation.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. doi: 10.1037//0033-2909.105.2.309

Sheldrake, R. (1998). Experimenter effects in scientific research: How widely are they neglected? *Journal of Scientific Exploration*, *12*, 73–78.

Shulman. R. (1938). An experiment in ESP with sounds as stimuli. *Journal of Parapsychology*, *2*, 322–325.

Smith, M. D. (2003). The role of the experimenter in parapsychological research. *Journal of Consciousness Studies*, *10*, 69–84.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—and vice versa. *Journal of the American Statistical Association, 54*. Retrieved from http://www.jstor.org/discover/10.2307/2282137?uid=3739560&uid=2&uid=4&uid=3739256&sid=21104078912057

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of outcome of statistical test on the decision to publish and vice versa. *American Statistician*. Retrieved from http://www.jstor.org/discover/10.2307/2684823?uid=3739560&uid=2&uid=4&uid=3739256&sid=21104078912057

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., & Di Risio, 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., & DiRisio, L. (2012). Meta-analysis of ESP Studies, 1987–2010: Assessing the success of the forced-choice design in parapsychology. *Journal of Parapsychology*, *76*, 243–273.

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

Taddonio, J. L. (1976). The relationship of experimenter expectancy to performance on ESP tasks. *Journal of Parapsychology*, *40*, 107–114.

Tressoldi, P. E. (2011). Extraordinary claims require extraordinary evidence: The case of non-local perception. *Frontiers in Quantitative Psychology and Measurement*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114207/. doi: 10.3389/fpsyg.2011.00117

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology*. Retrieved from http://www.frontiersin.org/Psychology/editorialboard

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science*, *6*, 363–378. doi: 10.1214/ss/1177011577

Utts, J., Norris, M., Suess, E., & Johnson, W. (2010). The strength of evidence versus the power of belief: Are we all Bayesians? In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. Voorburg, The Netherlands: International Statistical Institute.

Utts, J., & Tressoldi, P. (2013). Methodological and statistical guidelines. *Parapsychological Association*. Retrieved from http://www.parapsych.org/blogs/patrizio/entry/49/2013/2/methodological_and_statistical.aspx

Watt, C. (2007). Research assistants or budding scientists? A review of 96 undergraduate student projects at the Koestler Parapsychology Unit. *Proceedings of Presented Papers: The Parapsychological Association 50th Annual Convention*, 130–141.

Watt, C., Hopkinson, A., & Fraser, C. (2006). "Psychic DMILS": Attempted remote facilitation of performance in an ESP game, and an exploration of gender differences. *Proceedings of Presented Papers: The Parapsychological Association 49th Annual Convention*, 226–236.

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

Watt, C., Wiseman, R., & Schlitz, M. (2002, October). Tacit information in remote staring research: The Wiseman-Schlitz interviews. *Paranormal Review*, *24*, 18–25.

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

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

Willin, M. J. (2005). *Music, witchcraft and the paranormal*. Ely, England: Melrose.

Wiseman, R. (2010a)."Heads I win, tails you lose": How parapsychologists nullify null results. *Skeptical Inquirer*, *34*(1), 36–39. Retrieved from http://www.csicop.org/si/show/heads_i_win_tails_you_loser_how_parapsychologists_nullify_null_results/

Wiseman, R. (2010b, March 16). *Heads I win, tails you lose: How parapsychologists nullify null results*. Lecture presented at Goldsmith University, London. Retrieved from http://vimeo.com/11653478

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology*, *61*, 197–207.

Willamette University
900 State St. C267
Salem, OR 97301, USA
baptistajohann@gmail.com

**University of Nebraska - Lincoln*
1400 R St.
Lincoln, NE 68588, USA
maanelid@yahoo.com

Abstracts in Other Languages

German

JENSEITS DES MÜNZENWURFS: EINE NACHPRÜFUNG
VON WISEMANS KRITIK AN DER PARAPSYCHOLOGIE

ZUSAMMENFASSUNG: Wir überprüfen die von Professor Richard Wiseman vorgebrachte Kritik an der Parapsychologie, die er 2010 in seinem im Skeptical Inquirer erschienenen Artikel Heads I Win, Tails You Lose; How Parapsychologists Nullify Null Results formuliert hat, wobei wir seine wichtigsten Behauptungen detailliert widerlegen. Einige der von uns vorgebrachten Analysen bestehen aus Folgendem: Wir vergleichen die Reproduzierbarkeit von Psi-Experimenten mit derjenigen von Experimenten aus benachbarten Mainstreamgebieten - sie sind gleichwertig. Indem wir sowohl theoretische als auch empirische Zugänge verwenden, weisen wir nach, dass im Ganzfeld der file-drawer-Effekt keine Rolle spielt. Fälle von angeblicher Unwirksamkeitserklärung von Zufallsresultaten seitens des Experimentators werden genau geprüft und kritisiert. Die Schlussfolgerungen der Metaanalyse von Milton und Wiseman werden aufgrund der Ergebnisse von Bem, Palmer und Broughton sowie unserer eigenen Resultate hinterfragt. Angebliche Absinkungseffekte in den ASW-Paradigmen von Ganzfeld und begrenzter Wahl werden getestet und zurückgewiesen. Abschließend werden Strategien für einen Fortschritt präsentiert, die den überzeugendsten Trends und Konsistenzen zugrundeliegen, die wir im vorliegenden Datenmaterial gefunden haben. Wir stellen eine Nachprüfung der Kritik an der Parapsychologie vor, wobei uns Wiseman als maßgebliches Beispiel dient; wir zeigen auf, bis zu welchem Grad die Literatur skeptische Behauptungen zum einen nicht unterstützt und wie sie andererseits Psi als Erklärung für die Daten am plausibelsten erscheinen läßt.

Spanish

MÁS ALLÁ DE LANZAR UNA MONEDA : UN EXAMEN
DE LA CRÍTICA A LA PARAPSICOLOGÍA DE WISEMAN

RESUMEN: Examinamos la crítica de la parapsicología ofrecida por el Profesor Richard Wiseman en su artículo de 2010 , Heads I Win, Tails You Lose; How Parapsychologists Nullify Null Results, publicada en Skeptical Inquirer, y ofrecemos refutaciones detalladas de sus principales argumentos . Algunos de los análisis que llevamos a cabo son: Comparamos la reproducibilidad de los experimentos psi con la reproducibilidad de los experimentos en campos convencionales semejantes, mostrando que son equivalentes. Utilizando enfoques tanto teóricos como empíricos demostramos que los efectos de archivo (filedrawer effect) no están presentes en el ganzfeld. Examinamos y criticamos los casos de supuesta anulación por el experimentador de resultados nulos. Las conclusiones del meta- análisis de Milton y Wiseman son criticadas en base a los resultados de Bem, Palmer, y Broughton , así como de nuestros propios resultados. Examinamos y rechazamos los efectos de disminución ostensible en los paradigmas ganzfeld y de elección forzada de ESP. Finalmente, presentamos estrategias de progreso de acuerdo con las tendencias más atractivas y consistentes que hemos encontrado en los datos contemporáneos. Presentamos un análisis de la crítica en la parapsicología, con Wiseman como el ejemplo principal, que muestra el grado en el que la literatura no apoya las aseveraciones escépticas, así como la forma en que parece apoyar a psi como la explicación más plausible de los datos.

French

AU-DELA DU PILE OU FACE : EXAMEN DES CRITIQUES DE LA PARAPSYCHOLOGIE PAR WISEMAN

RESUME : Nous examinons la critique de la parapsychologie proposée par le Professeur Richard Wiseman dans article de 2010, Heads I Win, Tails You Lose; How Parapsychologists Nullify Null Results, publié dans le Skeptical Inquirer, et nous proposons une réponse détaillée à ses principales critiques. Parmi les analyses que nous avons conduites se trouvent les suivantes : nous avons comparé la reproductibilité des expérimentations psi à la reproductibilité des expérimentations dans d'autres champs de recherche mainstream, en découvrant qu'elles étaient équivalentes. En utilisant des approches à la fois théorique et empirique, nous avons démontré que les effets de fonds de tiroir ne sont pas présents dans le ganzfeld. Des cas de supposée nullification de résultats nuls par des expérimentations sont examinés et critiqués. Les conclusions de la méta-analyse de Milton et Wiseman sont confrontées aux résultats de Bem, Palmer et Broughton, ainsi qu'à nos propres résultats. D'apparents effets de déclin dans les paradigmes du ganzfeld et de la perception extra-sensorielle à choix forcée sont testés et rejetés. Finalement, nous présentons des stratégies pour progresser en nous basant sur les tendances les plus intéressants et cohérents découvertes dans notre base de données. Nous présentons un examen de la critique de la parapsychologie en prenant Wiseman comme exemple, montrant à quel point la littérature scientifique échoue à soutenir les revendications des sceptiques et à quel point elle fait de l'hypothèse psi l'explication la plus plausible des données.