CAN PARAPSYCHOLOGY MOVE BEYOND THE CONTROVERSIES OF RETROSPECTIVE META-ANALYSES?

By J. E. Kennedy

ABSTRACT: Retrospective meta-analyses are post hoc analyses that have not been effective at resolving scientific controversies, particularly when based on substantially underpowered experiments. Evaluations of moderating factors, including study flaws, small-study effects, and other sources of heterogeneity, do not neutralize confounding as in a well-designed experiment and cannot fully compensate for weaknesses in the original experiments. A group of well-designed experiments with adequate power and reliable results is needed for convincing evidence for a controversial effect. The widely recommended standard for experimental research is adequate power to obtain significant results on at least 80% of confirmatory experiments. Meta-analyses in parapsychology typically have found that 20% to 33% of studies with good methodology obtained significant results. Power analysis during experimental design is needed to achieve much better replication rates. Meta-analyses of RNG studies have consistently found that *z* value does not increase with samples size—which is contrary to statistical theory and has been and will be interpreted as an indication of methodological problems. This anomalous property and other sources of heterogeneity for parapsychological results must be addressed. Challenging topics such as experimenter effects, goal-oriented psi, and capricious psi-missing can no longer be ignored in research syntheses.

Keywords: power analysis, meta-analysis, experimenter effects, heterogeneity, synthesis-generated evidence

The field of parapsychology remains highly controversial and has not obtained the degree of acceptance and support that is needed. For the past 25 years, meta-analyses have been the foundation for the debates about the evidence for psi. This article focuses on the questions why have the meta-analyses been controversial and what can be done to move beyond these controversies?

Although the issues described here manifest in meta-analyses, the discussion covers much more than meta-analyses. Some of the key issues originate with the methodology and findings in the original experiments and must be addressed by appropriate new experiments. Also, alternative strategies for research synthesis may avoid some of the controversies associated with meta-analysis. Most of the final recommendations here do not involve meta-analysis.

The topics covered can be categorized as (a) intrinsic limitations of meta-analysis, (b) unfortunate experimental practices in parapsychological research, (c) problematic properties of the experimental findings in parapsychology, and (d) unfortunate meta-analysis practices in parapsychology. The combination of these factors has made parapsychological meta-analyses controversial. These categories interact, which requires that the same or similar topics are sometimes discussed under multiple categories.

This article does not attempt to comprehensively discuss all aspects of every issue. Some of the topics are controversial. The purpose here is to describe enough of the differing opinions to indicate practices that are not convincing if challenged.

Intrinsic Limitations of Meta-Analyses

The advent of meta-analysis in parapsychology in the 1980s was greeted with great enthusiasm. Small studies could be integrated to provide quantitative evidence for an effect and to evaluate potential moderating factors. Rosenthal (1986) and Utts (1986, 1991) argued that effect size was a more appropriate measure of replication than statistical significance. The usual practice of ignoring power analysis when designing experiments appeared to have good justification. Large studies were not needed. Meta-analysis was considered to provide the definitive evaluation of a line of research and to provide compelling evidence for psi. Broughton (1991) described meta-analysis as a "controversy killer."

However, this early optimism was not realized in practice. After noting cases when meta-analysis has been applied to controversial topics in psychology, Ferguson and Heene (2012) recently commented:

> [W]e have seldom seen a meta-analysis resolve a controversial debate in a field. … [W]e observe that the notion that meta-analyses are arbiters of data-driven debates does not appear to hold true. … [M]eta-analyses may be used in such debates to essentially confound the process of replication and falsification. … [F]ocusing on the average effect size may be used to, in effect, brush the issue of failed replication under the theoretical rug …. (p. 558).

The controversial debates noted in the article did not include parapsychology, but the comments aptly describe the experience with meta-analysis in parapsychology.

The limitations of meta-analyses were also apparent in medical research. Inconsistent or contradictory conclusions had been reached in different meta-analyses of the same database (Bailar, 1997). The statistical book most frequently used at a pharmaceutical company I recently worked with said the following:

> Our inclusion of [meta-analysis] in a chapter on exploratory analyses is an indication of our belief that the importance of meta-analysis lies mainly in exploration, not confirmation. In settling therapeutic issues, a meta-analysis is a poor substitute for one large well-conducted trial. In particular, the expectation that a meta-analysis will be done does not justify designing studies that are too small to detect realistic differences with adequate power. (Green, Benedetti, & Crowley, 2003, p. 231)

Ioannidis (2005) reached similar conclusions after developing methods for quantitative comparison of the "positive predictive value" PPV for different research methods. For adequately powered randomized experiments with little bias, he estimated the PPV to be .85. For meta-analyses of underpowered studies, the estimated PPV was .41, about half the PPV for a well-designed experiment. Similarly, evidence in pharmaceutical research is based on well-conducted experiments with adequate statistical power and reliable hypothesis tests (Food and Drug Administration, 1998). Retrospective meta-analyses cannot substitute for these well-designed, adequately powered experiments.

Several factors contribute to the limitations of meta-analyses. Some of the limitations have been discussed previously in parapsychological writings (Kennedy, 2004; Murray, 2011).

Many Choices for Post Hoc Analyses

Retrospective meta-analysis is a form of post hoc analysis. Like other types of post hoc analyses, meta-analysis involves many methodological decisions, including about statistical methods, study selection criteria, data trimming, data transformations, study quality ratings, and moderating factors. Many decisions do not have clear right and wrong answers. Different choices can result in different outcomes—which causes ambiguity and opportunity for selecting a preferred outcome.

The effects of different methodological decisions can be striking. For example, Bösch, Steinkamp, and Boller (2006a) describe some methodological differences between two meta-analyses by Radin on PK with electronic random number generators (RNGs; Radin, 1997; Radin & Nelson, 1989). The second meta-analysis reported an overall effect size that was much larger than in the previous analysis.

> The increase has two sources. First, Radin removed the 258 PEAR laboratory studies included in the first meta-analysis (without discussing why), and second, he presented simple mean values instead of weighted means as presented 10 years earlier. The use of simple mean values in meta-analyses is generally discredited … because it does not reflect the more accurate estimates of effect size provided by larger studies. In the case of the data in Radin's book, the difference between computing an overall effect size using mean values and using weighted mean values is dramatic. The removal of the PEAR laboratory studies effectively increased the impact of other small studies that had very large effect sizes. (Bösch, Steinkamp, and Boller, 2006a, p. 501)

The wide range of possible outcomes is also indicated by the RNG meta-analysis reported by Bösch, Steinkamp, and Boller (2006a). The overall outcome could be either significantly positive or significantly

negative (psi-missing), depending on whether a fixed-effects or random-effects model was used and whether the three largest studies were excluded as outliers.

As another example, the meta-analysis by Milton and Wiseman (1999) did not find significant evidence for psi, and was widely criticized for using cutoff criteria that excluded a highly significant study by Dalton (1997a). The subsequent meta-analysis by Storm, Tressoldi, and Di Risio (2010a) considered the Dalton study an outlier and excluded it from the analyses. However, other studies that had similar or larger effect sizes were not excluded.

For a typical retrospective meta-analysis, critics of the findings usually can easily find methodological decisions to challenge. These debates derive from the post hoc nature of meta-analysis and the associated potential for bias. The result has been endless controversies about meta-analysis methodology and findings in parapsychology (Bösch, Steinkamp, & Boller, 2006a, 2006b; Hyman, 2010; Radin, Nelson, Dobyns, & Houtkooper, 2006; Schmeidler & Edge, 1999; Storm, 2000; Storm, Tressoldi, & Di Risio, 2010a, 2010b). Like other types of post hoc analyses, meta-analysis can have value, but has limited effectiveness for resolving scientific controversies.

The Observational Nature of Moderating Factors

The evaluation of moderating factors in meta-analysis is observational or correlational analysis that does not neutralize confounding factors as in a well-designed experiment (Cooper & Hedges, 2009). For example, an evaluation of experimenter differences in a meta-analysis will typically be confounded by the experimenters' testing different pools of participants in different studies. Any differences could be due to the participants rather than the experimenters. Convincing evidence for experimenter differences must be based on planned experimental comparisons with different experimenters using the same pool of participants and same testing methods. In properly designed experiments, the independent or predictor variables are manipulated to eliminate confounding. Observational data do not have these experimental controls for confounding.

Meta-analysis methodologists now recognize that convincing conclusions about causality can come only from properly designed experiments, that is, *study-generated evidence* rather than *synthesis-generated evidence* (e.g., Cooper & Hedges, 2009). Synthesis-generated evidence "help[s] ensure that the next wave of primary research is sent off in the most illuminating direction" (Cooper & Hedges, 2009, p. 564).

The observational nature of moderating factors in meta-analyses is a significant limitation. For example, an evaluation of experimental flaws is not compelling evidence about the actual effects of the flaws. These correlational analyses are synthesis-generated evidence that cannot fully compensate for poor methodology in the original experiments. Confounding can cancel or dilute a real effect as well as artificially produce an effect. Carpenter and Palmer (1998) described a detailed example of apparent confounding in a meta-analysis in parapsychology.

Heterogeneity

When heterogeneous effect sizes are found in meta-analyses, "it is unclear whether the various research findings represent a common underlying phenomenon" (Wood & Eagly, 2009, p. 459). The usual recommendation is to identify the sources of heterogeneity and use appropriate subgroups or models for the moderating factors (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, Hedges, & Valentine, 2009). If the sources of heterogeneity cannot be fully identified, random-effects analyses are generally considered more appropriate.

Moderating factors have a central role when working with heterogeneous data. However, the observational nature of moderating factors brings into focus the limitations of drawing conclusions from heterogeneous data. Meta-analyses should incorporate known moderating factors, but the possibility of unknown confounding factors is always present with observational data.

Most of the issues discussed in later sections of this paper typically manifest in a meta-analysis as heterogeneous effect sizes and cannot be convincingly resolved by the evaluation of moderating variables. Appropriately designed experimental studies are required.

Biases From Small Studies

In medical research it has become well established that substantially underpowered studies are prone to elevated effects known as *small-study effects* (Egger, Smith, Schneider, & Minder, 1997; Sterne, Egger, & Smith, 2001; Sterne, Gavaghan, & Egger, 2000). Most, but not all, factors causing small-study effects are forms of methodological bias. Substantially underpowered studies tend to be exploratory and therefore subject to practices like multiple analyses, post hoc analyses, data selection, and optional stopping. Investigators and journals tend to not report nonsignificant analyses from underpowered studies because they are inconclusive. However, significant findings are likely to be reported. Small studies usually have fewer experimental personnel and less formal procedures than large studies, and therefore have higher potential for some type of experimenter effect, including fraud. In addition, small studies are also more likely to have a limited or selected range of subjects. In general, the greater effort to plan and conduct experiments with larger sample sizes usually results in better methodology and greater likelihood of full publication with peer review.

The need to evaluate and handle small-study effects in meta-analyses is widely recognized. Various methods have been proposed (Begg, 1994; Egger, Smith, Schneider, & Minder, 1997; Ioannidis & Trikalinos, 2007b; Sutton, 2009).

However, these methods provide synthesis-generated evidence and are unreliable if the effect sizes are heterogeneous, if all the studies are small, or if there are few studies (Ioannidis and Trikalinos, 2007a, 2007b; Sutton, 2009). Meta-analyses dominated by underpowered studies can be expected to be controversial. Large studies are the frame of reference for evaluating small-study effects. The reliability and strength of the evidence increase as the number of studies with adequate power increases.

Conclusions

The original hope that retrospective meta-analysis would provide convincing evidence for a controversial effect has not been realized in practice. After describing the post hoc, observational aspects of meta-analysis, Cooper and Hedges (2009) emphasized that "a research synthesis should never be considered a replacement for new primary research" (p. 564).

The most convincing evidence for a meta-analysis occurs when all included experiments are well designed with adequate power and obtain reliable effects. If some experiments are underpowered or have flaws, or the effects are heterogeneous, observational synthesis-generated analyses are used to attempt to compensate for the weaknesses in the original experiments. The conclusions are much less convincing if weak studies are a significant part of the evidence. A group of well-designed experiments with adequate power and reliable results is needed for convincing evidence for a controversial effect.

Unfortunate Experimental Practices

The use of power analysis to design adequately powered experiments is essential for controversial areas of research. Unfortunately, power analysis has rarely been used in designing parapsychological experiments. The overly optimistic reliance on meta-analyses apparently resulted in many experimenters ignoring power analysis. However, experimenters increasingly appear to recognize that this strategy is not effective for a controversial area of research.

An underpowered study is a biased form of research because it cannot provide basic evidence that the experimental hypothesis is false. An underpowered study is likely to produce nonsignificant results and creates ambiguity about whether nonsignificant results are due to the lack of power or to the experimental hypothesis being false. A significant result is interpreted as evidence supporting the experimental hypothesis, but a nonsignificant result is inconclusive. However, for an adequately powered study, a nonsignificant result is evidence that the experimental hypothesis is false.

Underpowered studies can also generate false positive results. Bakker, van Dijik, and Wicherts (2012) discussed the common practice in academic psychology of conducting a series of small, underpowered studies rather than a large study. This strategy generates many analyses that are not treated as multiple

analyses and therefore enhance false positive results. Simulation studies verified that "executing multiple small and underpowered studies represents the optimal strategy for individual players to generate a *p* value of less than .05 [by capitalizing on chance]" (p. 547). This practice has also been common in parapsychology, and some examples for ganzfeld research are noted below.

In my two decades of work in medical research, the majority of researchers preferred to evaluate the evidence for an effect by focusing on adequately powered studies and discounting small, underpowered studies. This strategy has also been recommended in psychology (Kraemer, Gardner, Brooks, & Yesavage, 1998). The primary value of small studies is to justify the effort for conducting larger studies and to develop parameters for the studies.

The minimum power typically recommended in both behavioral and medical research is .8 (e.g., Cohen, 1988, p. 56; Food and Drug Administration, 1998, p. 22). If an effect is real, at least 80% of properly designed confirmatory studies should obtain significant outcomes. This degree of replication provides convincing evidence that the experimenters understand and control the phenomena being investigated. A power of .90 or .95 is preferable when possible.

As shown in Table 1, the majority of meta-analyses in parapsychology have found that 20% to 33% of the individual studies obtained statistically significant results. For meta-analyses with 30 or more studies with good methodology by a variety of experimenters, the rate of successful replication is between 20% and 33%. Cases with a smaller number of studies and/or possible methodological problems sometimes have replication rates outside of this range.

By the usual methodological standards recommended for experimental research, there have been no well-designed ganzfeld experiments. Based on available data, Rosenthal (1986), Utts (1991), and Dalton (1997b) described 33% as the expected hit rate for a typical ganzfeld experiment where 25% is expected by chance. With this hit rate, a sample size of 201 is needed to have a .8 probability of obtaining a .05 result one-tailed.[1] No existing ganzfeld experiments were preplanned with that sample size. The median sample size in recent studies was 40 trials, which has a power of .22.[2]

Conducting a series of small studies has been a common ganzfeld research strategy. Research programs with more than 200 total trials from a series of small studies have been reported by Bem and Honorton (1994), Kanthamani and Broughton (1994), and Broughton and Alexander (1997). The combined results were significant for Bem and Honorton, but not for the other two research programs. The total numbers of trials and studies in these experimental programs apparently were not preplanned, which introduces a post hoc component for the combined analyses.

The replication rates of 20% to 33% appear to apply to all types of studies. The RNG studies are the largest experimental database and have some very large experiments, yet 25% or less were significant.

Conclusions

The most convincing evidence for an effect is driven by well-designed confirmatory experiments with adequate power to reliably obtain significant results. For the past 25 years, the efforts to develop evidence for psi have focused more on retrospective meta-analyses of underpowered studies than on conducting well-designed confirmatory experiments.

Problematic Properties of Psi Experiments

Small-Study Effects

Meta-analyses in parapsychology have often found characteristics expected for small-study effects. In the first ganzfeld meta-analysis, Hyman (1985) reported that the small studies had unexpectedly large

1 The sample size of 201 is from the free G*Power program (Faul, Erdfelder, Lang, & Buchner, 2007; 2012) that uses an exact power calculation. A power calculation using an arcsin transformation (SWOG, n.d.) gives a sample size of 199. Most power calculators use a normal approximation, which gives a sample size of 192.

2 The power calculation method is more important for small sample sizes such as 40. The exact method gives a power of .22, the arcsin method gives .30, and the normal approximation gives .33. Exact power plotted against sample size can have a saw-tooth or jagged form for small sample sizes.

effects. Similar findings were reported for later ganzfeld studies in the Bem and Honorton (1994) meta-analysis, and in the RNG meta-analyses by Radin, May, and Thomson (1985), Radin and Nelson (2003), and Bösch, Steinkamp, and Boller (2006a). The negative relationships between scoring rate and study size reported for the early ESP studies (Nash, 1989) also indicated larger effects for small studies. In the absence of a convincing alternative explanation, cautious scientists will assume these effects result from methodological biases.

Table 1
Properties of Meta-Analyses in Parapsychology

Meta-Analysis		Significant studies		*N* trials in the studies		*z* & √*N* trials correlation	
		Counts	Percent	Median	Max	*r*	*p*
Ganzfeld							
Honorton, 1985		12/28	43%	28	100	.13	.25
Bem & Honorton, 1994	- All	2/11	18%	35	50	-.12	.64
	- Unbiased	1/10	10%	36	50	-.08	.59
Milton & Wiseman, 1999		6/30	20%	40	100	.16	.20
Storm, Tressoldi, & Di Risio, 2010a	- All	10/30	33%	40	138	.48	.004
	- Trimmed	9/29	31%	40	138	.38	.02
Combine 3 from above	- All	18/71	25%	40	138	.34	.002
(1994, 1999, 2010a)	- Trimmed	16/69	23%	40	138	.28	.009
Bem, Palmer & Broughton, 2001a, 2001b		9/29	31%	40	128	.34	.04
Standard ganzfeld studies							
RNG							
Radin & Nelson, 1989	- All	152/597	25%				
	- Trimmed	490					
Radin & Nelson, 2003		515			$3.9x10^8$	-.02	.36
Bösch, Steinkamp, Boller, 2006a	- All	83/380	22%	8596	$>10^9$	-.14	.006
	- Trimmed	83/377	22%	8039	$>10^8$	-.02	.66
Other							
Honorton & Ferrari, 1989.	- All	92/309	30%	1194	$3.0x10^5$	.16	.003
Forced-choice precognition	- Trimmed	62/248	25%				
Radin & Ferrari, 1991.	- All	65/148	44%	5500	$2.4x10^5$		
PK with dice	- Unbiased	23/69	33%				
	- Trimmed	59					
Lawrence, 1998. Sheep-Goat		18/73	24%	5750	$5x10^4$	.22	.08
Storm, Tressoldi, &	- Noise reduction	3/16	19%	44	120	.40	.06
Di Risio, 2010a	- Free response	3/21	14%	54	937	-.20	.80
	- Trimmed	0/14	0%	76	937	.11	.35

The meta-analyses shown here have been prominent and have a reasonable number and quality of studies. Other meta-analyses with few studies or with studies with questionable quality were not considered suitable for the analyses here.

The counts of significant studies are all positive one-tailed results. *N* Trials are the number of trials or individual random events in the experiment. The correlations for *z* and √*N* are one-tailed Pearson correlations unless noted otherwise. According to standard statistical theory, *z* should increase linearly with √*N*.

For Bem and Honorton (1994), the original report focused on 10 studies and excluded one that had potential confounding by response bias.

For Milton and Wiseman (1999) two data values are corrected per Bem, Palmer, and Broughton (2001b).

Radin and Nelson (2003) is an update of Radin and Nelson (1989) with additional studies. The number of studies is reported as smaller (515 compared to 597) because they collapsed the 258 PEAR studies into one data point for the 2003 analysis. No rationale was given for that decision. Both meta-analyses included nonintentional and nonhuman studies and some studies with pseudo-RNGs. The *z* and √*N* correlation is two-tailed and was reported in the original meta-analysis.

For Bösch, Steinkamp, and Boller (2006a), the *z* and √*N* correlations are two-tailed and were reported in the original meta-analysis.

For Honorton and Ferarri (1989), the correlation is apparently a Pearson correlation between *z* and *N* (not √*N*) and was reported in the original meta-analysis.

Sample Size and *z*

A simple and intuitive way to evaluate the consistency of effects for a set of studies is to examine the relationship between *z* and samples size. According to standard statistical theory, the *z* value is expected to increase linearly with the square root of the sample size. This is the rationale for using *z* divided by $\sqrt{N}$ as an effect size measure that is independent of sample size. This relationship is also the basis for power analysis. Correlations between *z* and $\sqrt{N}$ for meta-analyses in parapsychology are given in Table 1. Of course, these are synthesis-generated evidence that primarily motivate future research.

The most striking finding is that the RNG studies clearly show a complete absence of positive correlation between *z* and $\sqrt{N}$. The fact that *z* is independent of sample size has been consistently recognized from the earliest meta-analyses of RNG studies (Bösch, Steinkamp, & Boller, 2006a; Radin, May, & Thomson, 1985; Radin & Nelson, 2003; Radin, Nelson, Dobyns, & Houtkooper, 2006).

The matter is less clear for the ganzfeld studies. Consistent with the RNG studies, the first three ganzfeld meta-analyses in Table 1 did not find significantly positive ($p \leq .1$) correlations. However, the meta-analysis by Storm, Tressoldi, and Di Risio (2010a) did show a significantly positive correlation between *z* and $\sqrt{N}$. The positive relationship is also apparent when those data are combined with data from earlier meta-analyses as shown in Table 1.

One important question is whether the first three ganzfeld meta-analyses had sufficient statistical power to demonstrate the correlation between *z* and $\sqrt{N}$. I performed simulations that assumed the overall hit rate for each meta-analysis applied for all studies in the meta-analysis (the assumptions for a fixed-effects model). The power for generating a correlation with $p \leq .1$ was evaluated, which is generally considered an appropriate significance level for this type of analysis. Data were generated simulating 2,000 meta-analyses for the studies in each meta-analysis. For the Honorton (1985) meta-analysis (hit rate 36.8%, based on equivalent hit rate for the *z* value for studies with chance not 25%), the correlation between *z* and $\sqrt{N}$ had a .83 power of detecting a correlation at the $p \leq .1$ level. A correlation as small or smaller than the observed correlation ($r = .13$) occurred on only 5% of the simulations. For the Bem and Honorton (1994) meta-analysis (hit rate 32.2%), the power for the correlation was only .33. For Milton and Wiseman (1999; hit rate 27.6% from Milton, 1999, p. 313), the power was only .24. For Storm, Tressoldi, and Di Risio (2010a; hit rate 32.2%), the power was .77.

These results indicate that a significant correlation would be expected on the Honorton (1985) and Storm, Tressoldi, and Di Risio meta-analyses, but would not be expected for the Bem and Honorton meta-analyses and for the Milton and Wiseman meta-analysis. The lack of correlation for Honorton (1985) confirms Hyman's (1985) point that the small studies had larger effects than would be expected. Honorton (1985) argued that this result is due to differences in the experimental conditions and psychological factors rather than to methodological problems.

The meta-analysis of forced-choice precognition experiments by Honorton and Ferrari (1989) reported a significantly positive correlation, but these findings are questionable. These studies had very diverse experimental methods and extreme heterogeneity of results. The findings are questionable until the sources of heterogeneity are better understood. For example, the studies with RNGs may have different properties than the studies with cards, and these differences could confound the analyses. A favorable correlation was also reported for the meta-analysis of sheep-goat studies by Lawrence (1998), but an evaluation of the heterogeneity of the data was not reported. The other meta-analyses in Table 1 have few studies and varying results.

The issues discussed in this section can be convincingly resolved by demonstrating that future experiments with adequate power consistently obtain significant results. However, the usual methods for designing experiments do not apply if *z* values are unrelated to sample size. Unless the causes for this characteristic of the data are understood and controlled, RNG research will continue to have the following properties:

1. Power analysis cannot be used to design experiments. Larger sample sizes are not more likely to produce significant results. The replication rate may be limited to about 25%.

2. Standard statistical methods such as *t* tests, binomial tests, and ANOVA assume that each trial or subject is independent of the other trials or subjects and that power analysis is applicable. However, if the *z* value for an experiment is unrelated to sample size, these assumptions are violated in a way that makes the usual interpretations invalid.
3. The rationale for the usual measures of effect size in meta-analysis breaks down if *z* is independent of sample size. This makes the usual meta-analysis interpretations invalid.
4. Larger effect sizes for small studies are an established symptom of methodological problems and occur when *z* is independent of sample size. The majority of cautious scientists will find methodological problems to be the most plausible explanation for these results (e.g., Bösch, Steinkamp, & Boller, 2006a).

Several factors may have a role in these anomalous properties.

Experimenter Effects

Experimenter effects are one of the greatest challenges for parapsychology and should have a central place in any discussion of evidence for psi. Prominent experimenter differences have been recognized throughout the history of parapsychological research, and many experiments have established study-generated evidence for experimenter effects (see Kennedy & Taddonio, 1976; Palmer, 1997; Rao, 2011, pp. 170–197; White, 1976, for reviews). Most experimenters have often found nonsignificant outcomes on their experiments. However, a few experimenters have reported significant results on almost every experiment. The cause of the experimenter differences remains a matter of debate. The tendency for skeptics to obtain nonsignificant results is well known, but whether skepticism is more a cause or result of nonsignificant outcomes has not been resolved.

Experimenter misconduct is one possible factor contributing to the differences. Misconduct by an experimenter has occurred many times in parapsychology and is a constant threat (Kennedy, 2013). Experimenter misconduct includes biased analysis and reporting as well as fraudulently manipulating data. For example, an experimenter may present exploratory or post hoc analyses in a way that appears to be planned analyses. Or, an experimenter may conduct multiple hypothesis tests on an experiment but obtain a significant outcome on only one test, and then report it without mentioning that it was selected from multiple analyses.

Experimenter misconduct can be greatly reduced with prospective registration of studies, multiple-experimenter designs, and sharing data for independent analyses (Kennedy, 2013). I expect that these practices would significantly reduce the success rate for some experimenters.

Experimenter effects can cause apparent declines of effects. The initial findings for a line of research are usually from highly successful experimenters with 80% or more of the studies reported as significant. For example, the first six ganzfeld experiments by Honorton were all reported as significant (Honorton, 1977). Similarly, Schmidt (1973) reported that eight of the first nine studies he conducted with RNGs were significant. When other experimenters conduct replications, their rates of success are lower. The replication rates apparently decline to about 20% to 33% and then drift within this range as shown in Table 1.

Goal-oriented psi experimenter effects. The possibility that an experimenter influences the experimental outcome using psi is the most challenging form of experimenter effect. There is much evidence supporting this hypothesis (Kennedy & Taddonio, 1976; Palmer, 1997; Rao, 2011; White, 1976), including consistent evidence that successful experimenters are also successful subjects. Goal-oriented psi experimenter effects are conceptually the simplest form of experimenter psi, and the model most consistent with the RNG database.

Experimenters could obviously use psi to influence the outcomes of their experiments. PK research is based on the assumptions that psi is guided by the motivations of a person and can influence the outcome of a random process. Experimenters typically have high motivation for their experimental outcomes and the experiments are random processes.

Goal-oriented psi experimenter effects view the entire experiment as one random event with the probability of a hit of .05. From this perspective there is no difference between a person who wants to get a six on a die throw and a person who wants to get a significant outcome on an experiment. Both cases have motivation for the outcome of a random process.

Reviews of psi experiments have concluded that the complexity of the random process does not matter for a PK effect (Kennedy, 1978; Schmidt, 1987; Stanford, 1977). The term "goal-oriented psi" refers to the idea that PK depends on the desired outcome and is independent of the complexity of the random process. Thus, the fact that conducting an experiment is a more complex process than throwing a die would be irrelevant for a PK effect.

One of the main predictions of goal-oriented psi experimenter effects is that *z* will be unrelated to sample size on experiments (Kennedy, 1994, 1995). The larger sample sizes increase the complexity of the process but do not alter the goal for the outcome of the experiment. The RNG meta-analyses support the hypothesis of goal-oriented psi experimenter effects.

One major implication of goal-oriented psi experimenter effects is that process-oriented research is not meaningful. Successful experimenters can obtain whatever outcomes they want.

Studies of majority-vote processes also support the hypothesis of goal-oriented psi experimenter effects (Kennedy, 1995). According to standard statistical theory, majority-vote, or multiple psi efforts for one target, could be used to enhance the reliability of psi results. However, this standard method for increasing the reliability of a signal in noise is based on the same assumptions as increasing sample size in experiments. The experimental findings consistently indicate that very different outcomes are obtained on majority-vote experiments depending on the motivations and intentions of the experimenter (Kennedy, 1995). The pattern of results is consistent with efficient goal-oriented psi experimenter effects and is not consistent with the use of majority-vote processes to enhance psi accuracy.

The hypothesis of goal-oriented psi experimenter effects that are produced with the least possible occurrence of psi is consistent with the majority-vote results and with *z* being unrelated to sample size.

Overly Optimistic Assumptions. The assumptions for psi may be inconsistent with the assumptions for experimental research. Psi is assumed (a) to be guided by mental motivations, intentions, and needs, and (b) to produce effects that are not constrained by known physical parameters. However, experimental methods are based on physical parameters such as sample size, blinding, and randomization. The expectation that psi conforms to the physical parameters of experiments may be another case of overly optimistic hope. Goal-oriented psi experimenter effects could make sample size irrelevant. ESP appears to make blinding in an experiment impossible. PK by an experimenter could influence the participants' responses as well as the random events in an experiment. The ganzfeld procedure could make a participant more susceptible to psi influence by the experimenter rather than facilitate psi by the participant.

Most discussions of psi experiments appear to be based on the implicit assumption that the experiments are somehow miraculously immune to psi by the experimenter. Detached reflection may reveal that this assumption is conspicuously inconsistent with both the basic assumptions about the nature of psi and the experience with experiments. Rao (2011, p. 184) recently observed that "the psi experimenter effect remains a deeply disgusting predicament, which few researchers were and are willing to confront."

Capricious Psi-Missing

The unpredictable tendency for psi effects to be significantly opposite from the intended effect has been reported throughout the history of experimental parapsychology. Rao (1965) described the "bidirectionality" of psi that "shifts the mode of psi response from hitting to missing in a rather capricious manner" (p. 245). He described this characteristic as preventing the useful application of psi.

The RNG meta-analyses have found evidence for psi-missing, including significant missing on some large experiments. Psi-missing may contribute to the unexpected properties of the RNG studies. Meta-analyses typically are based on one-tailed analyses and will likely need to be adapted for use on at least some lines of research in parapsychology.

When psi-missing occurs at the level of the experimental outcome, two-tailed analyses and existing methodology are applicable. However, when psi-missing occurs within an experiment as described by Rao (1965, 2011, pp. 149–169), experimental design is much more challenging and the issues of post hoc and multiple analyses may be very difficult to overcome.

Several psi investigators have proposed that sporadic occurrences of psi are possible, but sustained results that provide convincing evidence and useful applications are not possible (reviewed in Kennedy, 2003). Capricious psi-missing and the associated failure to produce compellingly reliable results after many decades of research inspired this hypothesis. This hypothesis aptly summarizes the state of psi research, but requires an expanded view of scientific methodology.

Conclusions

The anomalous properties of psi experiments must be confronted and understood if progress is to be made in parapsychology. Ignoring the problematic properties in hopes that they will not be noticed or will go away is not a viable strategy for success. The ambiguities from underpowered studies tend to obscure the properties of psi experiments.

The lack of power analysis in parapsychology could be due in part to researchers recognizing that it is not applicable for psi. When I began working at the Institute for Parapsychology in 1974, the lab lore was that large studies were not more likely to obtain significant results and were very possibly less likely to be significant. I did an informal literature review at that time and found the evidence consistent with the lab lore, but, of course, there were too many confounding factors for convincing conclusions. I soon formed the impression that psi-mediated experimenter effects were a dominant factor for experiments with unselected participants, and I still find that hypothesis most consistent with the overall data. If these impressions are correct, experimental research that ignores these properties of psi will never be convincing or make scientific progress.

Unfortunate Meta-Analysis Practices in Parapsychology

Heterogeneity

As noted above, the standard recommendations for handling heterogeneity in a meta-analysis are to use appropriate subgroups or models for the moderating factors and use random-effects analyses if the moderating factors cannot be identified. However, the usual practice in parapsychology has been to trim (remove) data points that make the distribution heterogeneous and then to proceed assuming fixed-effects. The likelihood that this strategy ignores or distorts basic properties of the phenomena has rarely been considered. Among the many problems with this approach is that the trimmed studies have often been the largest studies, the ones that should have received the most attention. The meta-analysis by Honorton and Ferrari (1989) that found extreme heterogeneity and handled it by trimming 20 percent of the data is a conspicuous example of methodology that should be avoided. The only confident conclusion from that meta-analysis is that extreme heterogeneity was found.

Honorton, Radin, and others have sometimes argued that psi effects are intrinsically heterogeneous due to psychological factors (e.g., Honorton, 1985; Radin, Nelson, Dobyns, & Houtkooper, 2006). Unfortunately, most meta-analyses in parapsychology have not used methods that handle this intrinsic variability. For example, experimenter effects have not received the deserved attention in meta-analyses. Rosenthal (1986) noted that for the early ganzfeld studies the "investigators differed significantly and importantly in the average magnitude of the effects they obtained" (p. 327). However, many later meta-analyses in parapsychology have not evaluated experimenter differences.

Established sources of heterogeneity should be considered even if the test for heterogeneity is not significant. These tests have low power when the number or the sample sizes of the studies are small (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Also, exclusion of large studies because they are inconsistent with small studies is a dubious practice that should be minimized and very carefully justified if done. To the extent possible, the sources of heterogeneity should be identified and included in the analyses.

Stouffer's *Z*

The foundation for meta-analysis is the quantitative evaluation of the consistency of effects in different studies. Meta-analyses in parapsychology are widely discussed as evidence for replicable psi effects. Effect size is the basis for evaluating consistency of effects and replication. The standard methods to evaluate statistical significance for meta-analyses use weighted effect size measures that incorporate the consistency of effects and the greater reliability of larger studies.

However, most parapsychological meta-analyses have determined statistical significance using Stouffer's *Z*, which is based on *p* values rather than effect sizes and does not provide inferences about consistency or replication. Stouffer's *Z* tends to obscure inconsistency among studies and is particularly vulnerable to biases from small-study effects because it gives equal weight to all studies without regard for sample size.

Stouffer's *Z* is recommended only when effect sizes cannot be obtained or when a researcher is searching for any evidence of an effect without inferences about replication or effect size (Borenstein, Hedges, Higgins, & Rothstein, 2009, pp. 325–330). In the latter case, the basic purposes and methodology for meta-analysis are not applicable, and convincing evidence for a controversial effect cannot be expected. Stouffer's *Z* might be useful in situations when the results are basically convincing without statistical analyses—for example, well-designed, adequately powered experiments with heterogeneous but mostly significant outcomes. It might also be useful if the experimental results are goal-oriented psi experimenter effects that make sample size and effect size irrelevant.

Recommendations

The following recommendations appear to be essential if parapsychology is to progress beyond the current state of controversy.

Improve Experimental Methodology

The first recommendation is to improve experimental methodology. These practices are essential when findings will be professionally challenged. Key points include:

1. Experimental design must include appropriate sample sizes to obtain reliable results. Low replication rates and the limitations of underpowered studies do not provide convincing evidence for a controversial effect.
2. Experiments should be prospectively registered to eliminate various potential biases, as is standard procedure for areas of medical research (Kennedy, 2013). The Koestler Parapsychology Unit (2012) at the University of Edinburgh provides a study registry, as well as information about the development of other registries.
3. Experimental procedures should make it difficult for one experimenter to intentionally or unintentionally alter data (Kennedy, 2013).
4. Raw data should be made available for independent analyses. If biased data fishing is likely, an investigator may reasonably require that the recipient register the planned analyses, including corrections for multiple analyses, prior to receiving the data (Kennedy, 2013).

Parapsychological research has generally followed the methodological practices of academic psychology and manifests methodological problems that are common in academic psychology. The need to overcome these problems is increasingly recognized—as indicated by a special issue of *Perspectives on Psychological Science* (Pashler & Wagonmakers, 2012, available online) that should be required reading for all social and behavioral scientists. These problems have also been discussed in other recent articles (e.g., Laws, 2013; Simmons, Nelson, & Simonsohn, 2011).

In general, the field of parapsychology would benefit from looking to medical research for experimental methodology rather than to academic psychology. Study registration is a well-established

practice in medical research, whereas psychological researchers are just beginning to develop study registries. My experience in medical research has been that formal power analysis is standard procedure, and the limitations of a series of substantially underpowered studies are widely recognized. My experience in parapsychology has been that formal power analyses and concern about the limitations of underpowered studies have been rare. The references in the previous paragraph strongly suggest that experimental psychology and parapsychology are similar on these matters.

Understand Anomalous Properties of RNG Experiments

The second recommendation is to attempt to understand the independence of *z* and sample size for the RNG studies. One option is to use meta-analysis to explore moderating variables. A more powerful option is to build simulation models that have the characteristics of the RNG database. In addition to methodological flaws and experimenter and participant differences, challenging effects like goal-oriented psi and capricious psi-missing will likely have a role. The database of RNG studies could be posted on the internet to encourage investigation. Of course, the ultimate goal of these post hoc explorations is to develop predictions for prospective research. Goal-oriented psi may be investigated with experiments using majority-vote processes.

Based on the history of psychical research, I predict that as larger studies are conducted and more diverse experimenters are involved, the ganzfeld and other lines of research will increasingly manifest the anomalous properties found for the database of RNG studies.

Appropriate Methods for Research Synthesis

The third recommendation is to use appropriate methods for parapsychological research synthesis. Unexplained heterogeneity is common and psychological factors are assumed to dominate the experimental results. The experimenters' motivations and goals often appear to be the most important factor for highly successful experimenters. Methods for research synthesis need to consider the variety of moderating factors. A random-effects model appears to be generally appropriate for formal meta-analyses in parapsychology and should always be reported, even if as a supplemental analysis.

Research syntheses that focus on well-designed confirmatory studies provide the strongest evidence for an effect. The simple proportion of adequately powered well-designed studies that obtained significant results can provide strong, robust evidence for an effect. In addition, a *best-evidence synthesis* (Slavin, 1986, 1995) that utilizes quantitative methods but limits the database to the currently best available studies is an increasingly recognized alternative to meta-analysis. The study selection criteria usually include a minimum sample size. Best-evidence syntheses also describe the strengths and weaknesses of the individual studies more than in a typical meta-analysis. Of course, best-evidence synthesis, like retrospective meta-analysis, is a form of post hoc analysis and involves decisions that can be biased.

Novel methods may need to be developed. The available data suggest that for RNG studies *z* might be a more appropriate measure of effect size than $z/\sqrt{N}$. This change would fundamentally alter the basic assumptions and methodology for analyzing data. Careful consideration of the implications is needed. The possibility that this methodological change applies primarily for certain experimenters should also be considered.

References

Bailar, J. C. (1997). The promise and problems of meta-analysis. *New England Journal of Medicine*, *337*, 559–561.

Bakker, M., van Dijik, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. Retrieved from http://pps.sagepub.com/content/7/6/543.full.pdf+html

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis and meta-analysis* (1st ed.) (pp. 399–409). New York: Sage.

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

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

Bem, D. J., Palmer, J., & Broughton, R. S. (2001b). Errata. *Journal of Parapsychology, 65*, 427–428.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. , & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

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

Bösch, H., Steinkamp, E., & Boller, E. (2006b). In the eye of the beholder: Reply to Wilson and Shadish (2006) and Radin, Nelson, Dobyns, and Houtkooper (2006). *Psychological Bulletin,* 132, 533–537.

Broughton, R. S. (1991). *Parapsychology: The controversial science*. New York: Ballantine Books.

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

Carpenter, J. C., & Palmer, J. (1998). Comments on the extraversion-ESP meta-analysis by Honorton, Ferrari, & Bem. *Journal of Parapsychology, 62*, 277–282.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, H., & Hedges, L. V. (2009). Potential and limitations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 561–572). New York: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Sage.

Dalton, K. (1997a). Exploring the links: Creativity and psi in the ganzfeld. *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention*, 119–134.

Dalton, K. (1997b). Is there a formula for success in the ganzfeld? Observations on predictors of psi-ganzfeld performance. *European Journal of Parapsychology, 13*, 71–82.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal, 315*, 629–634. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2127453/pdf/9310563.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2012). G*Power 3. Software retrieved from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561. Retrieved from http://pps.sagepub.com/content/7/6/555.full.pdf+html

Food and Drug Administration. (1998). *Guidance for industry: E9 statistical principles for clinical trials*. Retrieved from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073137.pdf

Green, S., Benedetti, J., & Crowley, J. (2003). *Clinical trials in oncology* (2nd ed.). New York: Chapman & Hall/CRC.

Honorton, C. (1977). Psi and internal attention states. In B. B. Wolman (Ed.), *Handbook of parapsychology* (pp.435–472). New York: Van Nostrand Reinhold.

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

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

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

Hyman, R. (2010). Meta-analysis that conceals more than it reveals: Comment on Storm et al. (2010). *Psychological Bulletin, 136*, 486–490.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, 0696–0701. Retrieved from http://cmajopen.com/content/176/8/1091.full.pdf+html

Ioannidis, J. P. A., & Trikalinos, T. A. (2007a). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, *176*, 1091–1096. Retrieved from http://cmajopen.com/content/176/8/1091.full.pdf+html

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

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. Retrieved from http://jeksite.org/others/hk_1994_pa.pdf.

Kennedy, J. E. (1978). The role of task complexity in PK: A review. *Journal of Parapsychology, 42*, 89–122. Retrieved from http://jeksite.org/psi/jp78.pdf.

Kennedy, J. E. (1994). Exploring the limits of science and beyond: Research strategy and status. *Journal of Parapsychology, 58*, 63–82.

Kennedy, J. E. (1995). Methods for investigating goal-oriented psi. *Journal of Parapsychology, 59*, 47–62.

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

Kennedy, J. E. (2004). A proposal and challenge for proponents and skeptics of psi. *Journal of Parapsychology, 68*, 157–167.

Kennedy, J. E. (2013). Experimenter misconduct in parapsychology: Analysis manipulation and fraud. Retrieved from http://jeksite.org/psi/misconduct.pdf.

Kennedy, J. E., & Taddonio, J. L. (1976). Experimenter effects in parapsychological research. *Journal of Parapsychology*, *40*, 1–33. Retrieved from http://jeksite.org/psi/jp76.pdf.

Koestler Parapsychology Unit. (2012). Registry for Parapsychological Experiments. Retrieved from http://www.koestler-parapsychology.psy.ed.ac.uk/TrialRegistry.html

Kraemer, H. C., Gardner C., Brooks J. O., & Yesavage J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, *3*, 23–31.

Lawrence, T. R. (1998). Gathering in the sheep and goats: A meta-analysis of forced-choice sheep-goat studies, 1947–1993 [Abstract]. In N. L. Zingrone, M. J. Schlitz, C. S. Alvarado, & J. Milton (Eds.), *Research in parapsychology 1993* (pp. 27–30). Lanham, MD: Scarecrow Press.

Laws, K. R. (2013). Negativland—a home for all findings in psychology. *BMC Psychology, 1*, 1–8. Retrieved from http://www.biomedcentral.com/content/pdf/2050-7283-1-2.pdf

Milton, J. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part I. Discussion paper and introduction to an electronic-mail discussion. *Journal of Parapsychology*, *63*, 309–333.

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

Murray, A. L. (2011). The validity of the meta-analytic method in addressing the issue of psi replicability. *Journal of Parapsychology*, *75*, 261–277.

Nash, C. (1989). Intra-experiment and intra-subject scoring declines in *Extrasensory Perception After Sixty Years*. *Journal of the Society for Psychical Research, 55*, 412–416.

Palmer, J. (1997). The challenge of experimenter psi. *European Journal of Parapsychology*, *13*, 110–125.

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530. Retrieved from http://pps.sagepub.com/content/7/6.toc

Radin, D. I. (1997). *The conscious universe*. San Francisco: HarperEdge.

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

Radin, D. I., May, E. C., & Thomson, M. J. (1985). *Psi experiments with random number generators: Meta-analysis Part 1*. Unpublished manuscript.

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

Radin, D. I., & Nelson, R. D. (2003). Research on mind-matter interactions (MMI): Individual intention. In W. B. Jonas & C. C. Crawford (Eds.), *Healing, intention and energy medicine: Research and clinical implications* (pp. 39–48). Edinburgh, UK: 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.

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

Rao, K. R. (2011). *Cognitive anomalies, consciousness, and yoga*. New Delhi, India: Centre for Studies in Civilization.

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

Schmeidler, G. R., & 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, H. (1973). PK tests with a high-speed random number generator. *Journal of Parapsychology, 37*, 105–118.

Schmidt, H. (1987). The strange properties of psychokinesis. *Journal of Scientific Exploration, 1*, 103–118. Retrieved from http://www.scientificexploration.org/journal/jse_01_2_schmidt.pdf

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. Retrieved from http://pss.sagepub.com/content/22/11/1359.full.pdf+html

Slavin, R. E. (1986). Best evidence synthesis: An intelligent alternative to meta-analysis and traditional reviews. *Educational Researcher, 9*, 5–11.

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology, 48*, 9–18.

Stanford, R. G. (1977). Experimental psychokinesis: A review from diverse perspectives. In B.B. Wolman (Ed.), *Handbook of parapsychology* (pp. 324–381). New York: Van Nostrand Reinhold.

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Systematic reviews of health care: Investigating and dealing with publication bias and other biases in meta-analysis. *British Medical Journal, 323*, 101–105. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120714/

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and presence in the literature. *Journal of Clinical Epidemiology, 53*, 1119–1129.

Storm, L. (2000). Research note: Replicable evidence of psi: A revision of Milton's (1999) meta-analysis of ganzfeld databases. *Journal of Parapsychology, 64*, 411–416.

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

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010b). A meta-analysis with nothing to hide: Reply to Hyman (2010). *Psychological Bulletin, 136*, 491–494.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L.V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 435–452). New York: Sage.

SWOG. (n.d.). One sample binomial. Online power calculator retrieved from http://www.swogstat.org/stat/public/one_binomial.htm

Utts, J. (1986). Successful replication versus statistical significance. *Journal of Parapsychology, 52*, 305–320.

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

White, R. A. (1976). The limits of experimenter influence of psi test results: Can any be set? *Journal of the American Society for Psychical Research, 70*, 333–369. Retrieved from http://www.aspr.com/limits.htm.

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 455–472). New York: Sage.

Broomfield, CO, USA
jek@jeksite.org
http://jeksite.org/psi.htm