THE VALIDITY OF THE META-ANALYTIC METHOD IN ADDRESSING THE ISSUE OF PSI REPLICABILITY

By Aja Louise Murray

ABSTRACT: Meta-analytic techniques are held in particularly high esteem in parapsychology owing to their important contribution to debates on the controversial issue of psi replicability. They are, however, associated with some serious limitations. The present paper evaluates the extent to which these limitations have represented a significant impediment to the resolution of replicability issues in psi research. It concludes that the subjectivity inherent in the execution of the technique and the interpretation of meta-analytic results has led to a situation whereby it has not been able to provide definitive results on the question of psi replicability.

Keywords: Meta-analysis, replication, psi, subjectivity

Meta-Analysis and Replication in Psi Research

Replication is critical in demonstrating that a given result is not due to chance or artifact (Lykken, 1968) and, indeed, most traditional philosophies of science list replicability as a requisite for scientific study (Attmanspacher & Jahn, 2003; Godfrey-Smith, 2003). Within psychology, much of the controversy surrounding both the existence of psi and parapsychology's scientific status has centred on a purported lack of repeatable results in psi research (Beloff, 1994; Irwin & Watt, 2007; Milton & Wiseman, 2001). Given this, it seems imperative that parapsychologists seek replicability of psi effects. Parapsychologists are acutely aware of this need and, historically, much energy has been devoted to this end (Utts, 1991).

Meta-analysis has played a prominent role in this goal: it has found application across a range of experimental domains in ESP (e.g., Bem & Honorton, 1994; Bem, Palmer, & Broughton, 2001; Haraldsson, 1993; Honorton, 1985; Honorton & Ferrari, 1989; Honorton et al., 1990; Honorton, Ferrari, & Bem, 1998; Hyman, 1985; Lawrence, 1993; Milton, 1997a; Milton & Wiseman, 1999; Radin, 2005; Sherwood & Roe, 2003; Stanford & Stein, 1994; Steinkamp, Milton, & Morris, 1998; Storm, 2000; Storm & Ertel, 2001; Storm, Tressoldi, & Di Riso, 2010) and PK (Bösch, Steinkamp, & Boller, 2006a; Braud & Schlitz, 1997; Radin, 1997; Radin & Ferrari, 1991; Radin & Nelson, 1989, 2003; Schmidt, Schneider, Utts, & Walach, 2004) research, and its results are held in high esteem (e.g., Palmer, 2003). Storm (2006), for example, describes meta-analysis as a "Godsend for parapsychologists" (p. 37) and one critic has suggested that the arguments for the consistency of ganzfeld results rest solely on meta-analytic evidence (Hyman, 2010). There is no doubt that meta-analysis has played a major

role in the ganzfeld debates (Palmer, 2003), and the importance of the technique in other experimental domains appears to be growing.

Given the widespread enthusiasm for meta-analysis, it is of critical importance to enquire as to the extent to which the technique yields valid and reliable evidence bearing on the psi replicability question. The present paper will describe some of the most pertinent limitations and advantages of meta-analysis in the context of psi research and evaluate the extent to which they have respectively undermined and enhanced the technique's contribution to addressing the issue of whether there is replicable evidence for psi.

Meta-analysis is used to obtain a quantitative synthesis of the individual (primary level) studies relevant to a given research question. To a first approximation, the enthusiasm for meta-analysis in addressing psi replicability would appear to be entirely justified. This is because the technique can both summarise the average size of an effect across multiple studies in a single index and provide a rich set of auxiliary statistics pertaining to effect size moderators, confidence intervals, consistency across studies, statistical significance, and indications of the likelihood of results being due to publication bias (Borenstein, Hedges, Higgin, & Rothstein, 2009; Palmer, 2003). Each of these, directly or indirectly, provides a means of evaluating replicability. Meta-analysis, therefore, seems to offer myriad riches when it comes to addressing the question of psi replicability. These sources of evidence are discussed in more detail below.

The most fundamental source of evidence for replicability offered by meta-analysis is a nontrivial effect size abstracted from several occasions of asking the same research question (Rosenthal, 1991). Were effects not replicable, the resulting abundance of null or chance negative results would act to decrease this combined effect size to a negligible magnitude. As random errors will cancel out with conglomeration, meta-analysis also overcomes the problem of noise and pseudofailure to replicate at the primary research level when studies are underpowered (Bayarri & Berger, 1991; Broughton, 1991; Rosenthal, 1986; Storm, 2006). Biases such as the precision-sample size or quality-effect size relations that may, at the primary level, obscure or give the illusion of replicability, can be partly eradicated by weighting studies by sample size or study quality (Borenstein et al., 2009; Storm, 2006). Combined z scores and p values that are used to infer statistical significance can be calculated on the same principles (Borenstein et al., 2009).

In a number of instances, the size of these main effects has favoured a psi research hypothesis (e.g., Radin & Nelson, 1989; Schmidt et al., 2004) but in other cases they have not (e.g., Milton & Wiseman, 1999). It is not sufficient, however, to rely exclusively on these indices as evidence for the replicability of psi effects. This is because replicability additionally implies consistency of results across studies. A significant main effect, however, can arise in the presence of marked heterogeneity and, conversely,

heterogeneity can mask what might otherwise represent a significant main effect (Borenstein et al., 2009; Bem et al., 2001). Meta-analysis addresses this problem by providing a measure of effect size heterogeneity such as *Q* (Laird & Mosteller, 1990) or the I^2 statistic (e.g., Cuijpers, Smit, Bohlmeijer, Hollon, & Andersson, 2010). These indices represent critical supplemental tests of replicability because without homogeneity any claim of replicable effects is undermined.

Even apparent evidence of replicability (sizeable main effect, nonsignificant heterogeneity) can be due to selection bias, but this can also be addressed within meta-analysis because it affords the opportunity to investigate possible publication bias and its influence. Publication bias is indicated by asymmetry of a funnel plot: a graphic representation with effect size on the X axis and sample size, variance, or standard error on the Y axis (Egger, Smith, Schneider, & Minder, 1997). This asymmetry can be quantified and used as the basis for a judgment as to the presence and extent of publication bias (e.g., Higgins & Green, 2008). Publication bias may also be indicated by an inverse or lack of correlation between study size and effect size (Bösch et al., 2006a). The extent to which publication bias has influenced meta-analytic main effects can be investigated through methods such as Orwin's (1983) failsafe *N*. This method is a modification of Rosenthal's (1979) failsafe *N* that assesses the number of unpublished studies required to bring the meta-analytic main effect to a specified level deemed to reflect an effect of no substantive importance. The larger this number, the smaller the potential impact of publication bias. Another option is Duval and Tweedie's "trim and fill" method that successively removes the most extreme small studies to yield a symmetric funnel plot and a corresponding unbiased effect size (Duval & Tweedie, 2000). The attenuation of variance is corrected by adding the original studies and their imputed mirror image back into the analysis (Duval & Tweedie, 2000). The larger the discrepancy between the original and corrected effect size, the greater the likely impact of publication bias. When publication bias and its influence are evident, confidence in meta-analytic main effects are undermined (Bösch et al., 2006a; Darlington & Hayes, 2000; Rosenthal, 1979, 1995).

Similarly, an apparently replicable psi effect may be nothing more than a replicable methodological artifact, but again, this can be addressed in meta-analysis using moderator analyses. In particular, moderation by methodological quality where poorer quality studies yield larger effect sizes has been taken as indicative of potentially artifactual results (Honorton, 1985; Palmer, 2003; Utts, 1991, 1993).

Subjectivity in Meta-Analysis

One might be tempted to conclude that, given that meta-analysis is comprehensive in its coverage of potential issues pertaining to replicability,

it should yield conclusions which can be accepted with a high degree of confidence. The reality of the situation in psi research, however, is that these sources of evidence are far from perfect, and this undermines the certainty of meta-analytic results. Meta-analyses are not automated, objective procedures: they are conducted by humans and, as such, are vulnerable to errors and cognitive biases. Errors may be less problematic as they are usually easily identifiable; for example, Radin, Nelson, Dobyns, and Houtkooper (2006) quickly identified that Bösch et al. (2006a) had omitted a large study from their meta-analysis of RNG studies. Cognitive biases, however, leave a less obvious trace. While there is little doubt that meta-analysis is more objective than the narrative review approach to assessing replicability through evidence synthesis (Krippner et al., 1993; Johnson & Eagly, 2000), there remain a number of subjective decision points and, thus, opportunities for the introduction of the effects of cognitive bias (Wanous, Sullivan, & Malinak, 1989). This includes defining and judging studies against inclusion criteria (Kennedy, 2004; Palmer, 2003), search strategies (Kennedy, 2004), coding studies (Glass, McGaw, & Smith, 1981; Milton, 1996; Steinkamp, 1998) and identifying and dealing with outliers (Wanous et al., 1989) or methodologically poor studies, including underpowered studies (Kraemer, Gardner, Brooks, & Yesavage, 1998; Rosenthal, 1991). Truly blinded coding of studies is difficult to implement in parapsychology (Steinkamp, 1998) because the field is small in size and it is difficult to set out coding criteria in advance of possessing knowledge of study outcome (Watt, 2005). As a result, some researchers choose to reject blinded coding, arguing that only naïve coders can be truly blinded (Schmidt et al. 2004). As coding requires a degree of familiarity with psi research methods, reliance on nonexperts may not be a viable option (Schmidt et al., 2004).

That the consequences of such subjective decisions are not mere theoretical possibilities is evidenced by the impact that they have on both main and auxiliary meta-analytic results in psi research. Several authors have noted that, in general, different meta-analytic procedures can lead to different outcomes (Bailar, 1997; Fishbain, Cutler, Rosomoff, & Rosomoff, 2000; Morris, 1991; Wanous et al., 1989) and different meta-analysts working with the same database can arrive at quite disparate conclusions (Nestoriuc, Kriston, & Rief, 2010; Watt, 2005). Within psi research, Milton (1997b) showed that stronger meta-analytic main effects could be obtained using sum of ranks rather than direct hits as the outcome variable in a database of free response ESP studies. Milton and Wiseman's (1999) decision to include nonstandard ganzfeld studies dramatically reduced the size of the effect, as standard and nonstandard procedures were found to differ to a statistically significant extent (Bem et al., 2001). Schmidt et al. (2004) compared the use of only good quality studies (a best evidence synthesis; Slavin, 1995) to simply weighting all studies by quality, and only in the latter case was there a significant main effect. Finally, Bösch et al. (2006a) treated the three largest studies in their RNG database as outliers; however, had they not done so,

they would have found their results to be in the opposite direction in their fixed effect model (Bösch et al., 2006a). Wilson and Shadish have (2006) questioned whether it was appropriate to treat these studies as outliers. In fact, the random and fixed effects models also differed by a statistically significant amount in Bösch et al. (2006a), but in the absence of a detailed understanding of psi effects and their distributional properties, there is no compelling reason to think that either model is more appropriate (Borenstein et al., 2009).

Assessments of heterogeneity, moderators, and publication bias may also be affected by the outcome of subjective decisions. Homogeneity may be contrived by removing outliers as, for example, Hyman (2010) has argued occurred in Storm et al.'s (2010) meta-analysis of ESP studies. Such practices are not uncommon in parapsychology and may be quite extreme (Delanoy, 1993). Radin and Nelson (1989), for example, report that up to 45% of studies may be removed for the sake of achieving homogeneity. With regard to moderation by study quality, Hyman (1985) and Honorton (1985) arrived at opposite conclusions despite analysing the same database of ganzfeld studies. Each reported an outcome consistent with their own theoretical disposition—Hyman, the critic, found a correlation between study quality and outcome whereas Honorton, the proponent, found no such correlation (Palmer, 2003). Steinkamp (1998) reports that the level of disagreement between coders analysing study quality in the Steinkamp et al. (1998) meta-analysis of clairvoyance and precognition sometimes reached as much as 66%. Subjective decisions can also lead to more or less conservative estimates of likelihood of the results being due to publication bias (Macaskill, Walter, & Irwig, 2001). For example, alternative failsafe *N* methods can lead to quite divergent estimates (Rothstein, Sutton, & Borenstein, 2005). Indeed, in Storm et al.'s (2010) meta-analysis, allowing for the possible presence of negative results in the file drawer made a substantial difference to estimates of the impact of publication bias. The authors estimated that for the ganzfeld studies in their database, using Rosenthal's (1995) fail safe *N*, 293 nonsignificant studies would need to be in the file drawer to bring their results to a nonsignificant level. This was compared to their estimate of 86 studies using Darlington & Hayes's (2000) method. Together, these observations imply that although procedures exist to minimise the influence of unreliable individual studies and selection biases on the meta-analytic main effects, their application entails a subjective judgment on the part of the researcher. This can lead and has led to markedly different results dependent on the outcome of this judgment in psi research.

Subjectivity also abounds in the interpretation of meta-analytic results, wherein different theoretical dispositions can again lead to quite divergent interpretations (Bösch et al., 2006b). In many cases it is not possible to arrive at a consensus as to whether a meta-analysis indicates replicability (Palmer, 2003). Part of the problem is that the different sources

of evidence described above have a tendency to conflict with one another. In Schmidt et al. (2004), for example, moderation by study quality called into question the extent to which their overall significant main effect constituted evidence for replicable DMILS. In Bösch et al. (2006a), confidence in the significant main effect in the RNG studies is undermined by significant heterogeneity and the likely presence of publication bias. There are no agreed upon standards for precisely what conditions must be met in order to conclude unequivocally that replicability is in evidence (Palmer, 2003). The weight that should be afforded to each source of evidence and which should take precedence when they are in discord is, therefore, largely up for debate (Palmer, 2003). An assessment is particularly difficult to make if this set of evidence is not reported in its entirety. Radin and Nelson (2003), for example, tested neither moderators nor heterogeneity in their meta-analysis of PK studies (Bösch et al., 2006a).

This same subjectivity is apparent in the interpretation of the meaning of individual results within a given meta-analysis. For example, the inverse relation between study size and effect size in Bösch et al. (2006a) could be interpreted either as evidence of publication bias or of psychological moderators of effect size, with smaller studies being more psi conducive (Radin et al. 2006). The effect size of the same study was also a source of disagreement, prompting debates about the extent to which it was so small as to be essentially meaningless (Jarrett, 2006; Wilson & Shadish, 2006).

Thus, despite the promises of meta-analysis, there remains a situation whereby some proponents, such as Radin (1997), view psi results as being as consistent as those in the physical sciences, while critics remain wholly unconvinced (Hyman, 2010). It would seem that critics and proponents will always be able to cite the limitations of meta-analysis: its mostly retrospective nature (Hyman, 2010); its dependence on the quality of primary level research (Nestoruic et al., 2010); subjectivity (Eysenck, 1994); selection biases (Noble, 2006); the "apples and oranges" problem (Glass et al., 1981); and its strictly quantitative, reductionist nature (Bösch et al., 2006a), as undermining positive and null results, respectively. Proponents and critics alike have always proven adept at explaining away such criticisms from the "opposition" (e.g., Kennedy, 2006). Such discourse highlights the fact that meta-analysis in psi rarely yields results simultaneously convincing to both critics and proponents. This lack of consensus can be attributed, at least in part, to the room for subjectivity allowed in the execution and interpretation of meta-analysis. The consequence of this is that it cannot be justifiably used as definitive evidence in support of either the proponent's or the critic's position.

Assuming that a consensus could be reached that a given meta-analysis contributed evidence in favour of psi replicability, this is no guarantee of future success. Meta-analyses themselves tend not to replicate well. Schmidt et al. (2004), for example, failed to replicate Schlitz and Braud's (1997) meta-analytic results in the DMILS domain. Even in the

ganzfeld domain, which typically yields some of the largest effect sizes (Hyman, 1991), there is a lack of consistency of results across different meta-analyses. Hyman (1985) and Honorton (1985) conducted meta-analyses of the ganzfeld studies—both finding statistically significant anomalous effects**.** Bem and Honorton (1994) analysed the studies conducted subsequent to these meta-analyses (10 auto-ganzfeld studies) and likewise found a statistically significant main effect. As Hyman (2010) notes, however, the significant result in the latter study was due solely to a subset of studies (those using dynamic targets), which calls into question whether this truly represents a successful replication. When Milton and Wiseman (1999) analysed ganzfeld studies conducted in the years following Bem & Honorton (1994), however, they found no statistically significant main effect at all, and even when it was updated and the overall effect brought up to a statistically significant level, the effect size was much smaller than that observed in the previous ganzfeld meta-analyses (Milton, 1999). This meta-analysis also failed to replicate two of the three moderators identified by Bem and Honorton (1994). The most recent meta-analysis of the domain (Storm et al., 2010) found a statistically significant overall effect for ganzfeld studies but the *z* scores behave differently in this new database compared to the older databases (Hyman, 2010). Specifically, whereas in the older database the *z* scores correlated negatively with the number of trials in an experiment, the relation was in the opposite direction for the Storm et al., (2010) analysis**.** Thus, there are reasons to doubt that this study represents a successful replication of the earlier ones (Hyman, 2010). Although it might be possible to explain the differences between meta-analytic results from the same domain—for example, the Milton and Wiseman (1999) study is argued to have included more studies with a greater emphasis on process-oriented rather than proof-oriented research (Bem et al., 2001)—the main point is that meta-analyses in psi, for whatever reason, may not themselves be replicable. An individual meta-analysis, thus, is unlikely to be an adequately reliable source of evidence that psi effects are, or indeed are not, replicable.

It may be that there are, in fact, more fundamental problems with applying meta-analysis to the question of whether there are replicable psi effects. Psi has been characterised as inherently elusive and inconsistent, being as it is, outside the normal rules of the physical universe (Hyman, 2010; Kennedy, 2003; Kennedy, 2004). Indeed, psi results often do not conform to the assumptions of standard statistical models: sample size may be unrelated to statistical significance (e.g., Radin & Nelson, 2000) or the two may be inversely related (e.g., Bem & Honorton, 1994; Steinkamp et al., 2002). If this is not just due to publication bias and is, in fact, a property of psi, then the meaningfulness of effect size and thus of meta-analytic results are seriously undermined (Kennedy, 2004). If this concern has a basis in reality, the use of meta-analysis is perhaps inappropriate in psi research in attempting to fit psi effects to the scientific model of replicability. That

being the case, however, proponents cannot simultaneously maintain that favourable results in meta-analysis constitute evidence for the replicability of psi effects.

Finally, it is worth noting that irrespective of whether meta-analysis represents a truly valid and reliable source of evidence for replicability, it can at least assist in improving the reliability and validity of the manner in which the replicability issue is addressed. Through moderator analyses or descriptive comparisons of different conditions it has been possible to identify putative psi conducive procedures, suggesting the conditions under which replicability is most likely to occur if it is to be found at all. Such psi conducive conditions identified by meta-analysis include certain experimenters (Rosenthal, 1986), participants who study a mental discipline (Milton & Wiseman, 1999), ganzfeld procedures rather than other noise reduction techniques or no noise reduction techniques (Storm et al., 2010) and standard rather than nonstandard ganzfeld procedures (Bem et al., 2001). Furthermore, the scrutiny under which studies are placed in analysing them for meta-analysis can identify sources of bias and methodological shortcomings. This occurred, for example, when Radin and Ferrari (1991) found evidence that dice throwing experiments were subject to a bias due to the differential weight of the die faces, or when Hyman and Honorton (1986) published guidelines for conducting future ganzfeld studies based on shortcomings identified while meta-analysing results from the domain. It can also inform future experimental designs by providing an estimate of expected effect size—an estimate which allows the experimenter to calculate the number of participants required to run an adequately powered experiment (Utts, 1991). Thus meta-analysis can assist in identifying and overcoming the factors that present barriers to replicability.[1]

Future Directions

Of course, improving the application of meta-analytic procedures in psi research in itself will also lead to addressing the question of psi replicability with greater reliability and validity. For example, mandatory preregistration of primary level studies and prospective meta-analyses may attenuate problems of optional stopping, post hoc analyses, missing data, unfalsifiability, and publication bias (Kennedy, 2004; Watt, 2005), which all promote uncertainty in the accuracy of meta-analytic results. Where it is not possible to have complete control over primary level studies, prespecification

1 A related point is the use of meta-analysis in process-oriented research. The present discussion has been limited to the merits of meta-analysis in evaluating psi replicability and, therefore, in proof-oriented research. The problems identified do not necessarily preclude meta-analysis as a potentially useful tool in exploratory process-oriented research. A separate discussion is necessary to address this related but also somewhat distinct issue given the diverging goals of proof- and process-oriented research (Irwin & Watt, 2007).

of meta-analyses can be beneficial. Caroline Watt and Richard Wiseman, for example, issued a call for the preregistration of all replication attempts of Bem's (2011) "feeling the future" study, stating in advance the cutoff date for study registration and completion for inclusion in their meta-analysis. In setting out the decisions that are the source of much of the contention in the interpretation of meta-analyses (inclusion and coding criteria, statistical models, inferential tests, etc.) with justification in advance of the research being conducted, greater objectivity and transparency can be achieved. Parapsychologists have already demonstrated their willingness to report their rationales for the decisions taken in the meta-analytic process and to participate in public discourse evaluating such decisions (e.g., Palmer, 2010). This has the dual benefit of both allowing the reader to make a fully informed judgement as to the appropriateness of any subjective decisions taken and promoting self-reflection on the part of the researcher. To do this in advance of conducting the meta-analysis, where possible, would represent a further improvement to practices.

Where meta-analysts in psi research are not afforded the luxury of preregistered individual studies, it would be prudent to incorporate the observations from mainstream science that inflated effect sizes tend to appear in primary level studies which are conducted earlier (Ioannidis, 2008), published in higher impact journals (Munafo, Stothart, & Flint, 2009), or have smaller sample sizes (Kraemer et al., 1998). In many psi meta-analyses, publication date and sample size are already being examined as effect size moderators (e.g., Schmidt et al. 2004; Steinkamp et al., 2002). Parapsychologists would also be justified in investigating the viability of making corrections for such effects.

Another suggestion for the improvement of meta-analytic investigations in psi research has been the use of Baysean meta-analysis to replace the dominant frequentist approaches (Dawson, 1991; Utts, 1991). The rationale for this suggestion is that, among other things, Baysean techniques are more explicit in the utilisation of prior knowledge in hypothesis testing (see Utts et al., 2010 for a review of the other potential advantages of Bayes in psi research). Baysean meta-analysis, however, may prove equally prone to the problems of subjectivity in fields such as parapsychology where the estimation of priors can elicit somewhat polarised responses dependent on theoretical disposition (Bem, Utts, & Johnson, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Moreover, the issue of whether and when Baysean approaches are more appropriate than their frequentist counterparts is one which has long been the subject of a debate that has transcended research domains and is by no means limited to psi research (McGrayne, 2011). The superiority of Baysean meta-analysis over frequentist methods is, therefore, not clear but does represent a reasonable line of enquiry.

Finally, it is interesting to note that irrespective of whether one views psi research as akin to, a control group for, as creditable as, or a

hindrance to mainstream psychology research, the points discussed in the present paper have implications that extend to psychology research more generally. As in parapsychology, the importance of replication for weeding out spurious results and establishing phenomena is widely acknowledged and discussed within mainstream psychology (e.g., Munafo & Flint, 2010). Yet these concerns are not always reflected in practice, as most findings in mainstream psychology are not subject to a replication attempt (Schmidt, 2009), and even when they are, imprecise definitions of what constitutes a replication can lead to pseudoreplication (Sullivan, 2007). Nor is meta-analysis in mainstream psychology free of many of the problems reviewed here (e.g., Bailar, 1997; Rothstein et al. 2005). Indeed, the issues discussed in the present paper are only one manifestation of the more general problem of the subjectivity in putative scientific practice (Longino, 1990; Kitcher, 2001). Other statistical techniques, other research questions, and other branches of human enquiry, including the natural sciences, are to a greater or lesser extent hindered by issues of subjectivity. Parapsychologists and other researchers alike should, thus, endeavour to maintain appropriate levels of scepticism regarding their own beliefs and practices. Of course, this idea is not new (e.g., Chamberlin, 1897), but the arguably disproportionate enthusiasm for meta-analysis in addressing the question of psi replicability is perhaps an example of the importance of keeping this in focus.

Conclusions

Meta-analysis provides invaluable evidence bearing on the question of whether there is replicable evidence for psi. But it also suffers from a number of limitations, perhaps the most problematic of which is subjectivity of procedures and interpretation. Given its limitations, definitive results are rarely attained and debates about psi replicability remain largely unresolved. The solution to this problem is not to discard meta-analytic results but to continue to make improvements to the technique, seeking ever more objective and stringent procedures. Although meta-analysis fails to always deliver definitive answers, it remains the closest approximation to a valid and reliable investigation of psi replicability currently available (Irwin & Watt, 2007).

References

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

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

Bayarri, M. J., & Berger, J. (1991). Comment. *Statistical Science, 6,* 379–382.

Beloff, J. (1994, August). *The sceptical position: Is it tenable?* Paper presented at the 37th Annual Convention of the Parapsychological Association, Amsterdam, The Netherlands.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425.

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

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

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Reply: Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology, 101,* 716–719.

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

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

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

Chamberlin, T. C. (1897). The method of multiple working hypotheses. *Journal of Geology, 5,* 837–848.

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *British Journal of Psychiatry, 196,* 173–178.

Darlington, R. B., & Hayes, A. F. (2000). Combining independent *p* values: Extensions of the Stouffer and binomial methods. *Psychological Methods, 5,* 496–451.

Dawson, R. (1991). Comment. *Statistical Science, 6,* 382–385.

Delanoy, D. L. (1993, August). *Experimental evidence suggestive of anomalous consciousness interactions.* Proceedings of the Second Gauss Symposium, Munich, Germany.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

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.

Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal, 309,* 789–792.

Fishbain, D., Cutler, R. B., Rosomoff, H. L., & Rosomoff, R. S. (2000). What is the quality of the implemented meta-analytic procedures in chronic pain treatment meta-analyses? *Clinical Journal of Pain, 16,* 73–85.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* London: Sage.

Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science.* Chicago: University of Chicago Press.

Haraldsson, E. (1993). Are religiosity and belief in the afterlife better predictors of ESP performance than belief in psychic phenomena? *Journal of Parapsychology, 57,* 259–273.

Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions.* Chichester, West Sussex: Wiley.

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

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

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, M., Derr, P., Schechter, E. I., & Ferrari, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology, 54,* 99–139.

Honorton, C., Ferrari, D. C., & Bem, D. J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Journal of Parapsychology, 62,* 255–276.

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

Hyman, R. (1991). Comment. *Statistical Science, 6,* 389–392.

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

Hyman, R., & Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology, 50,* 351–364.

Ioannidis, J. P. A. (2008). Why most true associations are inflated. *Epidemiology, 19,* 640–648.

Irwin, H. J., & Watt, C. A. (2007). *An introduction to parapsychology* (5th ed.). London: McFarland.

Jarrett, C. (2006). Evidence for psychokinesis, but is it meaningless? *The psychologist.* Retrieved from http://www.bps.org.uk/publications/thepsychologist/extras/pages$/2006-news/evidence-for-psychokinesis.cfm

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook*

of research methods in social and personality psychology (pp. 339–369). Cambridge, England: Cambridge University Press.

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

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

Kennedy, J. E. (2006). Letter on meta-analysis in parapsychology. *Journal of Parapsychology, 70,* 410–413.

Kitcher, P. (2001). *Science, truth and democracy.* Oxford, England: Oxford University Press.

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

Krippner, S., Braud, W., Child, I. L., Palmer, J., Rao, R., Schlitz, M., et al. (1993). Demonstration research and meta-analysis in parapsychology. *Journal of Parapsychology, 57,* 277–286.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6,* 5–30.

Lawrence, T. R. (1993). Gathering in the sheep and goats: A meta-analysis of forced-choice studies, 1947–1993. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention,* 75–86.

Longino, H. (1990*). Science as social knowledge: Values and objectivity in scientific enquiry.* Princeton, NJ: Princeton University Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Macaskill, P., Walter, A. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20,* 641–654.

McGraye, S. B. (2011). *The theory that would not die.* London: Yale University Press.

Milton, J. (1996). Establishing methodological guidelines for ESP studies: A questionnaire survey of experimenters' and critics' consensus. *Journal of Parapsychology, 60,* 289–334.

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

Milton, J. (1997b). A meta-analytic comparison of the sensitivity of direct hits and sums of ranks as outcome measures for free-response studies. *Journal of Parapsychology, 61,* 227–241.

Milton, J. (1999). Should the 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.

Milton, J., & Wiseman, R. (2001). Does psi exist? A reply to Storm & Ertel (2001). *Psychological Bulletin, 127,* 434–438.

Morris, R. L. (1991). Comment. *Statistical Science, 6,* 393–395.

Munafo, M. R., & Flint, J. (2010). How reliable are scientific findings? *The British Journal of Psychiatry, 197,* 257–258.

Munafo, M. R., Stothart, G., & Flint, J. (2009). Bias in genetic association studies and impact factor. *Molecular Psychiatry, 14,* 119–120.

Nestoriuc, Y., Kriston, L., & Rief, W. (2010). Meta-analysis as the core of evidence-based behavioral medicine: Tools and pitfalls of a statistical approach. *Current Opinion in Psychiatry, 23,* 145–150.

Noble, J. H. (2006). Meta-analysis: Methods, strengths, weaknesses and political uses. *Journal of Laboratory and Clinical Medicine, 147,* 7–19.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157–159.

Palmer, J. (2003). ESP in the ganzfeld—Analysis of a debate. *Journal of Consciousness Studies, 10,* 51–68.

Palmer, J. (2010). Meta-analysis of ESP ganzfeld studies [Letter to the editor]. *Skeptical Inquirer, 34,* 62.

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

Radin, D. I. (2005). The sense of being stared at: A preliminary meta-analysis. *Journal of Consciousness Studies, 12,* 95–100.

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

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. (2000). *Meta-analysis of mind-matter interaction experiments.* Unpublished manuscript, Boundary Institute, Los Altos, and Princeton Engineering Anomalies Research, Princeton University [cited in Kennedy, 2006].

Radin, D. I., & Nelson, R. D. (2003). Research in 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, Scotland: Churchill Livingstone.

Radin, D. I., Nelson, R. D., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on Bösch, Steinkamp, and Boller (2006) meta-analysis. *Psychological Bulletin, 132,* 529–532.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

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

Rosenthal, R. (1991). Meta-analysis: A review. *Psychosomatic Medicine, 53,* 247–271.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* Hoboken, NJ: Wiley.

Schlitz, M. J., & Braud, W. G. (1997). Distant intentionality and healing: Assessing the evidence. *Alternative Therapies in Health and Medicine, 3,* 62–73.

Schmidt, S., Schneider, R., Utts, J., & Walach, H. (2004). Distant intentionality and the feeling of being stared at. *British Journal of Psychology, 95,* 235–247.

Sherwood, S. J., & Roe, C. A. (2003). A review of dream ESP studies conducted since the Maimonides dream ESP studies. In J. Alcock, J. Burns, & A. Freeman (Eds.), *Psi wars: Getting to grips with the paranormal* (pp. 85–110). Thorverton, Exeter: Imprint Academic.

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

Stanford, R. G., & Stein, C. (1994). A meta-analysis of ESP studies contrasting hypnosis and a comparison condition. *Journal of Parapsychology, 58,* 235–269.

Steinkamp, F. (1998). A guide to independent coding in meta-analysis. *Proceedings of Presented Papers: The Parapsychological Association 41st Annual Convention,* 243–259.

Steinkamp, F., Milton, J., & Morris, R. (1998). A meta-analysis of forced-choice experiments comparing clairvoyance and precognition. *Journal of Parapsychology, 62,* 193–218.

Steinkamp, F., Boller, E., & Bösch, H. (2002). Experiments examining the possibility of human intention interactions with random number generators: A preliminary meta-analysis [Abstract]. *Journal of Parapsychology, 66,* 238–239.

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

Storm, L. (2006). Meta-analysis in parapsychology: I. The ganzfeld domain. *Australian Journal of Parapsychology, 6,* 35–53.

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 Riso, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136,* 471–485.

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science, 6,* 363–378.

Utts, J. (1993). Honorton the meta-analyst. *Journal of Parapsychology, 53,* 89–100.

Utts, J., Norris, M., Suess, E., & Johnson, W. (2010). The strength of evidence versus the power of belief: Are we all Bayseans? In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100,* 426–432.

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74,* 259–264.

Watt, C. (2005). Presidential address**:** Parapsychology's contribution to psychology: A view from the front line. *Journal of Parapsychology, 69,* 215–232.

Wilson, D. B., & Shadish, W. R. (2006). On blowing trumpets to the tulips: To prove or not to prove the null hypothesis—Comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin, 132,* 524–528.

University of Edinburgh
Department of Psychology
7 George Square
Edinburgh EH8 9JZ, UK
a.l.murray-2@sms.ed.ac.uk

Acknowledgments

Thank you to Caroline Watt and John Palmer for their input and to two anonymous reviewers for helpful comments on an earlier version of the manuscript.

Abstracts in Other Languages

French

LA VALIDITE DE LA METHODE META-ANALYTIQUE POUR RESOUDRE LA QUESTION DE LA REPLICABILITE DU PSI

RESUME: Les techniques méta-analytiques sont tenues en particulièrement haute estime en parapsychologie du fait de leurs importantes contributions aux débats sur la question controversée de la réplicabilité du psi. Elles sont néanmoins associées avec des limitations sérieuses. Le présent article évalue à quel point ces limitations ont représenté a frein significatif à la résolution des problèmes de réplicabilité dans la recherche psi. Il conclut que la subjectivité inhérente à

l'exécution de cette technique et à l'interprétation des résultats méta-analytiques mènent à une situation d'où il n'a pas été possible de fournir des réponses définitives à la question de la réplicabilité du psi.

Spanish

LA VALIDEZ DEL MÉTODO DE META-ANÁLISIS
PARA ABORDAR LA CUESTIÓN DE REPLICABILIDAD EN PSI

Resumen: Las técnicas de meta-análisis son tenidas en una estima especialmente elevada por la parapsicología, debido a su importante contribución a los debates sobre la controvertida cuestión de la replicabilidad psi. Empero, tienen serias limitaciones. Este trabajo evalúa en qué medida estas limitaciones han supuesto un obstáculo importante para la solución de los problemas de replicabilidad en la investigación en psi. La conclusión es que la subjetividad inherente en la ejecución de la técnica y la interpretación de los resultados de meta-análisis han llevado a una situación en la que no ha sido capaz de proporcionar resultados definitivos sobre la cuestión de la replicabilidad en psi.

German

DIE GÜLTIGKEIT DER META-ANALYTISCHEN METHODE
BEI DER BEHANDLUNG DER FRAGE NACH
DER WIEDERHOLBARKEIT VON PSI

ZUSAMMENFASSUNG: Meta-analytische Techniken werden in der Parapsychologie besonders hoch geschätzt, da sie einen wichtigen Beitrag zu den kontrovers geführten Debatten über die Replizierbarkeit von Psi liefern. Sie weisen jedoch einige ernstzunehmende Beschränkungen auf. Der vorliegende Beitrag wägt ab, inwieweit diese Beschränkungen ein ernstzunehmendes Hindernis bei der Lösung der Fragen nach der Wiederholbarkeit in der Psi-Forschung darstellen. Er kommt zum Schluss, dass die mit der Anwendung der Technik unvermeidlich gegebene Subjektivität und die Interpretation meta-analytischer Ergebnisse zu einer Situation geführt haben, in der es nicht möglich ist, endgültige Ergebnisse auf die Frage nach der Replizierbarkeit von Psi zu erwarten.