# Response to Howard (2018): Comments on 'A Meta-Reanalysis of Dream-ESP Studies'

Lance Storm[1], Adam J. Rock[2], Simon J. Sherwood[3], Patrizio E. Tressoldi[4], and Chris A. Roe[5]

[1]School of Psychology, University of Adelaide, Adelaide, Australia.
[2]School of Behavioural, Cognitive and Social Sciences, University of New England, Armidale NSW, Australia.
[3]Human Sciences Research Centre, University of Derby, Kedleston Road, Derby, UK.
[4]Dipartimento di Psicologia Generale, Università di Padova, Padua, Italy.
[5] School of Health and Society, University of Northampton, Northampton, UK.

*Summary*. Dream-ESP is a form of extra-sensory perception (ESP) in which a dreaming perceiver ostensibly gains information about a randomly selected target without using the normal sensory modalities or logical inference. We conducted a meta-analysis on dream-ESP studies (dating from 1966 to 2016), and found a number of significant effects indicating support for the ESP hypothesis (Storm et al., 2017). Howard (2018) critiqued our study, and found much weaker effects based on a re-analysis of our data, to which he applied inverse-variance weights to the study values. Although Howard replicated a number of our findings, his other findings can be challenged. We discuss meta-analytic approaches, including the controversial issues of publication bias and what to do with outliers, and we present some re-analyses.

*Keywords:* Dream-ESP, ESP, meta-analysis, psi

## 1. Introduction

In our original Dream-ESP meta-analysis (Storm, Sherwood, Roe, Tressoldi, Rock, & Di Risio, 2017), we defined Dream-ESP as a "form of extra-sensory perception (ESP) in which a dreaming perceiver ostensibly gains information about a randomly selected target without using the normal sensory modalities or logical inference" (p. 120). For the period 1966 to 2016, a homogeneous dataset of 50 studies yielded a mean z of 0.75 (*ES* = .20), with corresponding significant Stouffer Z = 5.32 ($p = 5.19 \times 10^{-8}$), suggesting that dream content can be used to identify target materials correctly and more often than would be expected by chance. The dataset was comprised of studies from two different periods:

- 14 studies from the 1960s to the mid-1970s—the Maimonides Dream Lab (MDL) studies;
- 36 studies from the early-1970s to 2016—'independent' (non-MDL) studies.

Although (a) the MDL dataset had a mean *ES* that was larger (.33) than the mean *ES* for the non-MDL studies (.14), the difference was not statistically significant; no statistically significant *ES* differences were found between (b) ESP mode (telepathy, clairvoyance, precognition); (c) REM and non-REM monitoring; (d) 'dynamic targets' (e.g., movie-film) and 'static targets' (e.g., photographs); (e) same-perceiver studies and different-perceiver studies; (f) same-agent studies and different-agent studies; (g) single-perceiver studies and multiple-perceiver studies; and (h) single-subject studies ($N = 1$) and multiple-perceiver studies ($N > 1$). We also found that significant improvements in the quality of the studies was not related to *ES*, but *ES* did decline over the five-decade period. In addition, we found no effect size difference between groups of authors. We also conducted a Bayesian analysis: the 95% Highest Density Interval of posterior probability (which indicates the most plausible 95% of the values in the posterior distribution) related to the *ES* ranged from 0.03 to 0.20 (mean *ES* = 0.12)—thus, the null hypothesis was rejected.

In a critique of our paper, Howard (2018) says he was "struck" by an "unusual feature" of our meta-analysis, and he suggested "the reported meta-analytic effect sizes are very large for psi research" (p. 224). We are surprised by this characterisation: The choice of phenomena from parapsychology that are taken as comparators by Howard seems peculiar, since they focus on claims that bear little relation to dream-ESP (including effects of mental intention on dice rolls). In fact, according to Cardeña (2018), effect sizes for psi dream studies are similar to (or smaller than) those for other free response psi effects, including remote viewing and ganzfeld experiments.

Using our original database (Appendix A in Storm et al., 2017, pp. 138-139), Howard (2018) replicates some of the effects we found in our meta-analysis, but he did not test for differences between 'same sender' and 'different sender', 'single-perceivers' and 'multiple-perceivers', 'dynamic targets' and 'static targets', or single-subject studies and multiple-perceiver studies; nor did he test whether improvements in the study quality were related to *ES*. However,

Corresponding address:
Lance Storm, PhD. School of Psychology, University of Adelaide, Australia.
Email: lance.storm@adelaide.edu.au

Submitted for publication: February 2019
Accepted for publication: February 2019

Howard does argue that additional modern meta-analytic methods would yield results that "may be notably different" (p. 224)—e.g., extra outlier identification methods, inverse-variance weighted meta-regressions, and within-group Cochran's Q (we note, however, that Cochran's Q is a measure of heterogeneity and tests the null hypothesis that all studies share a common effect size which has nothing to do with effect size comparisons). We do acknowledge in our paper that alternate analyses are possible and that our database may have imperfections.

Using our database, Howard performed an "inverse-variance weighted meta-analysis" (p. 224), and found a much smaller overall effect ($r$ = .07), where r is effectively the same as *ES* (i.e., $z/\sqrt{n}$). Inverse-variance weighted meta-analysis effectively reduces the strength of effects in smaller studies, of which there are a number in the Storm et al. (2017) database originating from the MDL. In fact, Howard claims "evidence that a significant relationship exists between effect size and sample size, suggesting that the prior results may have been primarily driven by large effects found in small-n studies" (p. 224). Actually, Howard found the effect (as a correlation between standard error and effect size) was only "marginally significant" for the whole sample, Kendall's $\tau$ = .165, $p$ = .058. He cites Borenstein (2005), who suggests that this test is "often underpowered", leading Howard to conclude that the "result suggests that a notable relationship exists between sample size and effect size in the dataset" (p. 228). However, we suggest it is not wise to build a case on a marginal effect; besides which, a correlation between *standard error* and effect size is not the same thing as a correlation between *sample size* and effect size. This preliminary finding is indicative of similar problems throughout Howard's paper.

In response to Howard (2018), we will address a number of key methodological issues, and then contrast Howard's findings with our past findings insofar as there are discrepancies, and we will include a few revisions. We find that the chief concerns centre around sample size, and how to deal with outliers appropriately.

## 2. Publication Bias

In interpreting findings derived from a database of published studies it is important to consider the effects of any publication bias, particularly the tendency for non-significant studies to be under-represented in the public record. In extreme cases, this can lead to Type I errors if substantial numbers of null or negative studies are unavailable. Running a test advocated by Darlington and Hayes (2000) to test for publication bias, Storm et al. (2017) found 110 unpublished studies must exist to reduce the database to non-significance. Howard (2018) reports that his re-analysis

> *found a similar result: the failsafe N was 93 for the analysis of all studies. However, the fail-safe N was noticeably smaller when the studies were separated by ESP mode: telepathy (0), clairvoyance (7), and precognition (0). This suggests that the current interpretations could be noticeably swayed by unpublished studies. (p. 225)*

This peculiar outcome might be of interest if there was indication that editors and/or authors had a preference for telepathy and precognition studies over clairvoyance studies. Otherwise, the fail-safe *N* is just fine, and the file-drawer argument has little support. However, and in the first place, the only justifiable reason to separate studies by modality is to test effect-size differences of the three modalities on the basis that they might involve fundamentally different psi processes, but the three effects did not differ: Storm et al. (2017, p. 127) did not find any differences in effect sizes by modality, and nor did Howard (2018, p. 226). We would argue that the rationale for testing for modality-dependent publication bias has to be theoretically grounded, and there does not appear to be a persuasive case for distinguishing among them, especially when in practice modalities are only differentiated operationally. Nonetheless, if Howard's estimate of 93 studies is legitimate, one can easily divide the 93 according to the proportions of each of the three ESP modes in Howard's Table 1 (p. 225), which has a total of 42 studies broken down by modality (two studies are not categorized). Hence, there would be 46 hypothesized studies for telepathy (i.e., 21/42 = 50%); 27 hypothesized studies for clairvoyance (i.e., 12/42 = 29%), and 20 hypothesized studies for precognition (i.e., 9/42 = 21%). These breakdowns suggest the number of hypothesized studies hidden away in the legendary file-drawer is too great to explain the observed data.

In any case, Howard (2018) has it the wrong way around in saying his "current interpretations could be noticeably swayed by unpublished studies", when in fact, "telepathy (0), … and precognition (0)" (p. 225), means there are *no hypothetical unpublished studies* for these two ESP modes, and thus no possible publication bias in the actual data for those modes. In effect, if we are going to talk about a publication bias, it seems the focus should not be on the whole sample (despite the significant Egger's test result indicating bias; p. 225), but should be on clairvoyance as that ESP mode produced a significant effect (see p. 226), possibly attributable to publication bias—quantitatively, an amount somewhere between 7 and 27 non-significant studies tucked away in file-drawers would reduce the effect to non-significance, but the other two ESP modes (telepathy and precognition) are off the hook. But then Howard disputes the finding of a significant effect for clairvoyance anyway, claiming the effect lacks "robustness" (p. 226). Note, however, that the effect size is still between .15 and .19.

Howard (2018) also argues that the inclusion of "imputed" (p. 226; i.e., alleged missing) studies would make the overall effect size non-significant. His Table 1 (p. 225) and Figure 1 (p. 226) show there are nine 'imputed' studies and the effect size is reduced from .072 to.049, which is no longer statistically significant when including the implied missing studies from the trim-and-fill analysis" (p. 225); the 95% CI changes from [.02, .12] to [-.01, .11], thus embracing zero. Considering the criticisms we have already raised here about publication bias and our disputing its likelihood, the act of adding 'imputed' studies is clearly nothing more than a theoretical exercise requiring justification.

## 3. Primary Analysis – Replications

Howard (2018) claims we made statistical decisions, including "reporting unweighted effects" (p. 224), that may have produced results that he regards as "inflated" (p. 224). He performed an "inverse-variance weighted meta-analysis" on our database. Two of Howard's analysis decisions seem to us to reflect *a posteriori* judgements that warrant comment here: rationale for inclusion and exclusion of studies in the final analysis; and weighting of effect sizes for included studies.

### 3.1. Selection of Studies for Inclusion

Guided by Cook's distance values, Howard excluded four significantly positive studies (#19, #25, #43, and #47), but only one significantly negative study (#2)—we had already classed two of these studies as outliers (#2 and #47) and removed them to create a homogeneous dataset—raising questions about whether the other three studies (#19, #25, #43) should have been removed. Howard states:

> *Eight methods were applied to find outliers and influential cases, but I primarily considered three in determining these studies: studentized deleted residuals, Cook's distance, and covariance rations (Viechtbauer & Cheung, 2010). (p. 225)*

Howard (2018) does not make clear his rationale for selecting three methods over the others available to him. Indeed, how to deal with outliers remains a controversial issue (see Orr, Sackett, & Dubois, 1991). There are other methods that could have been used (e.g., the 'Winsorized estimator', which is useful because it is relatively insensitive to outliers, but still gives a robust estimate of central tendency; Wilcox & Keselman, 2003). As an exercise, and starting with our homogeneous dataset ($N$ = 50), we set an outlier removal of 10% (which embraces the outliers in question), thus removing the lowest five and the highest five from the number set, and replacing them with the next closest entry. We found a Winsorized mean $ES$ = 0.199; mean $z$ = 0.73; and Stouffer $Z$ = 5.18 ($p = 1.11 \times 10^{-7}$). These values are comparable to our original values.

There are still other methods that could have been used by Howard (2018)—studies that are detected as outliers can be down-weighted using a Random Effect Variance Shift Outlier model (see Majd, Ghobadi, Baghban, Ahmadi, & Sajjadi, 2014). More broadly, removal of outliers and influential cases tends to be guided by what a given author may "recommend" (Viechtbauer & Cheung, 2010, p. 115), rather than hard-and-fast rules. For example, we note that Hunter and Schmidt (2004) even recommend against outlier analyses altogether, with the main reason being that "it is almost impossible to distinguish between large sampling errors and true outliers (i.e. actual erroneous data)" (p. 110). For transparency we note that if all outliers are retained then this would give a Winsorized mean $ES$ = 0.189; mean $z$ = 0.80; and Stouffer $Z$ = 5.80 ($p = 3.30 \times 10^{-9}$).

Howard (2018) states: "After their removal, the outlier analyses indicated that one large outlier still remained (Van de Castle, 1971)" (p. 225). He removed this study, and provided justifications for his decision in his 'Supplemental Material B'. We agree with Howard's concerns about the methodology of this study (see also, Sherwood & Roe, 2003, for similar doubts). The removal of this study (not as an outlier but for methodological reasons) gives (for a corrected $N$ = 49) a Winsorized mean $ES$ = 0.187; mean $z$ = 0.70; Stouffer $Z$ = 4.93 ($p = 4.11 \times 10^{-7}$). Or, for the original database (corrected $N$ = 51), a revised Winsorized mean $ES$ of 0.186; mean $z$ = 0.71; Stouffer $Z$ = 5.05 ($p = 2.21 \times 10^{-7}$).

We also note that Howard (2018) removed studies #11 and #48 because the procedure he used—"random-effects meta-analyses with inverse-variance weights"—cannot include studies "with sample sizes of two" (p. 225; italics added). We see no justification for the removal of studies merely for the sake of a test, and stress that these two studies are not outliers. One study (#11) has a zero effect and, more importantly, the other (#48) happens to have a positive value, with a medium $ES$ of .474. We also point out that, conventionally, in a random-effects model, one must justify the elimination of studies as outliers only on precise theoretical or methodological grounds, and not on the basis of their statistical distribution.

What we note throughout the various decisions made for one reason or another in Howard's paper is an unwarranted (apart from Van de Castle, 1971), systematic stripping down of the database of studies that mainly happen to demonstrate positive effects—studies which would ordinarily go towards supporting the psi hypothesis—whereas other methods that we have suggested are less severe on sensitive effects and/or do not require outlier removal—small wonder Howard arrives at the conclusion that "the results are more precarious than previously believed" (p. 228).

Finally, Howard (2018) refers to a change made to the database concerning a study by Watt (2014). We reported an $ES$ of 0.156 ($z$ = 2.20) in our original database because Watt (2014) had specifically stated that "data from any participants who did not complete four trials were discarded" (p. 105). But Watt and Valášek (2015) then adjusted this figure downwards by including those data that Watt initially excluded due to her exclusion criterion. Originally, we abided by Watt's criterion (see Storm et al., 2017, p. 132) but, to avoid complications, we opted for a lower figure in keeping with Howard's conservatism (NB: the above Winsorized calculations were made using these adjusted values). On a revised $N$ of 219 (up from 200), and three more hits (bringing the total hits up to 67), we see that Howard's effect size figure of .091, reported in his database, is incorrect, and should read 0.124̣ as reported in Watt and Valášek (p. 106; they actually reported a rounded-down figure of "0.12"). Also, Howard reported 247 trials, which should be 219. We spotted the source of the error in Howard's correspondence with the first author: "To calculate the Watt effect size, I added the additional 47 [*sic*] observations to the prior 200" (M. C. Howard, personal communication, November 5, 2018). The change from 0.156 to 0.124 results in a very slight drop in mean $ES$ for our database of .195 (down from .196, which we had rounded up to .20).

Howard's (2018) decision to use the lower figure for Watt (2014) is a further instance of *a posteriori* decision making regarding analysis strategy that in each case serves to reduce effect sizes. Would Watt's (2014) additional data have been included if they had served to inflate her reported effect size? We suspect not. Would the alternative methods for identifying and removing outliers referred to by Howard (p. 225), but not utilised in the analysis he presented, have produced similar reductions in effect size? We suspect not. Of course, it is possible to offer a plausible rationale for the choice that is made, but this does not protect against questionable research practices (cf. Steegen et al., 2016). No doubt another analyst with different motivations could have made arbitrary decisions that would serve to give inflated versions of the effect sizes we reported.

### 3.2. Weighting of Studies

Howard (2018) also argues that inverse-variance weighting is conducted because more credibility is attached to large studies (with smaller sampling variances) than smaller studies (with larger sampling variances). Since small studies are potentially more susceptible to publication bias, this can lead to small studies being viewed with suspicion, especially if they are significant. Howard noted that "no excessively

large-*n* studies overpowered the current results" (p. 225), but at the small-*n* end of the spectrum, he states: "Child and Krippner produced multiple outstanding effects in small-n studies, including effect sizes of .68, .72, and two of .94" (p. 228). However, it is important to recognise that small *n* studies may differ from large *n* studies in more respects than simply sample size, effect size, and susceptibility to sampling error. For example, many of the studies with $n < 15$ involved participants who were pre-selected based on prior performance, whereas no studies with $n > 15$ did so. Howard's strategy seems analogous to expressing doubt about goal-scoring in football matches because all observations involving intensive research with small samples of professional footballers have been rendered dubious by inappropriate generalisations from larger scale but less intensive studies of people who claim no particular footballing ability.

Likewise, we argue that small-*n* studies provide a more bespoke test of ostensibly psi-gifted participants than large-*n* studies, as more time and attention is given to single cases or small-*n* groups. J. B. Rhine (1948/1954) brought attention to this problem when he stated:

> *We destroy the phenomena in the very act of trying to demonstrate them. Evidently the tests themselves get in the way of the abilities they are designed to measure. (p. 161)*

Rhine's comment very readily suggests that 'conveyor-belt testing', consisting of multiple trials per participant, may be a disheartening process for participants (and/or experimenters). But even if there might be some statistical justification for authors "to conduct large-*n* studies of ESP in dreams" (Howard, 2018, p. 228), large-*n* studies could mean: (a) *few participants with many trials*; (b) *many participants with many trials*; or (c) *many participants with few trials*—Howard is not clear on which might be preferred, but we suggest they all carry the risk of decline effects to varying degrees.

Notwithstanding this problem, we query the premise that a number of small-*n* studies is *essentially* bad for parapsychology, even though meta-analysis can make such short work of them. Essentially, we can re-model the data to test this hypothesized alternative outcome, which would have the effect of giving grouped MDL studies smaller sampling variances. It should first be noted that we can use as a control, the mean *SE* (0.146) of the 12 MDL small-*n* studies (not including the Van de Castle study), each study ranging from two to eight trials. Based on year of publication, we split up the 12 small-*n* MDL studies, and formed three groups of 'moderate' size (which Howard sets at 15 to 99 trials). These three groups comprised:

- five studies with a total of 47 trials (one by Ullman, 1969; two by Ullman & Krippner, 1969; and two by Ullman, Krippner, & Feldstein, 1966);
- four studies with a total of 36 trials (Krippner, Honorton, & Ullman, 1972, 1973; Krippner, Ullman, & Honorton, 1971; Krippner, Honorton, Ullman, Masters, & Houston, 1971);
- three studies with a total of 18 trials (all three by Ullman, Krippner, & Vaughan, 1973).

The respective *SE*'s are 0.058, 0.066, and 0.094, which are (a) considerably smaller than the mean *SE* for the 12 small-*n* MDL studies of 0.146; and (b) in the vicinity of the non-MDL mean *SE* of 0.068 (minus the small-*n* studies). This crude demonstration suggests too much emphasis can be placed on the variance issue.

## 4. Discussion

For the rest of this response to Howard (2018), we will discuss some key issues about methodology and the implications and ramifications of Howard's findings, as well as our own. First, Howard carried out only a partial reanalysis of our meta-analysis, and did not clearly state his research objectives, other than to recalculate our effects using our Appendix A (Storm et al., 2017, pp. 138-139) to test the robustness of our findings. However, his reanalysis only considers our first four hypotheses, but not the remaining four. We make the very minor point that Howard's Results section, while it does differentiate between Primary and Additional analyses, might have been easier to follow if the specific planned analyses had been outlined in the Introduction.
Second, Howard (2018) stated:

> *Analyses were conducted to probe the effect of sample size on study results. Three groups were created that logically appeared in the database. The first included studies with a sample size over 99 (k = 7), the second included studies with a sample size of 15 to 99 (k = 13), and the third included studies with a sample size below 15 (k = 24). (p. 227)*

Howard regards these groups as "logically" appearing in the database, which is debatable because sub-divisions of this nature tend to be relatively arbitrary. There are a number of other (logical) ways to form groups. For example: we might accept the group with sample size over 99 ($k = 7$), but the second *could* include studies with a sample size of 10 to 99 ($k = 20$), and the third *could* include studies with a sample size $\leq 8$ ($k = 17$).

Third, Howard (2018) stated that,

> *The significant effect for clairvoyance studies was further investigated for robustness. The effect would no longer be statistically significant (r = .15 – 19, p > .05) when removing any of five studies (Dalton et al., 1999; Dalton et al., 2000; Kanthamani & Khilji, 1990; Roe et al., 2007; Sherwood et al., 2000). This suggests that this significant effect is somewhat precarious. (p. 226)*

Howard's strategy is to delete outlying scores/studies in order to find support for the null hypothesis, presumably on the basis that any one of these studies on its own is doing all the work of making clairvoyance look like a real effect. However, what is good for the goose, is good for the gander, and the arbitrary removal of a given study just because it has a *positive r* could equally be used to seek support for the alternative (psi) hypothesis for the two non-significant subsets (telepathy and precognition) by seeing what happens when non-significant (*negative r*) studies are removed one at a time.

Fourth, Howard (2018) stated that,

> *the 34 studies published in parapsychology journals produced an inverse-variance weighted effect size of .11 …, whereas the 10 studies published in non-parapsychology journals produced an inverse-variance weighted effect size of .02 (p. 228)*

Although the author only suggests "this difference may be significant", and he reports a "marginally significant result"

(p. 228), the implication may be that non-parapsychology journals are more methodologically stringent in their refereeing practices, and so these latter studies offer a more accurate estimate of the real effect size. Alternatively, this 'difference' might be thought to reflect the notion that orthodox ("non-parapsychology") journals are more inclined to publish psi studies that obtain null results (although parapsychology journals have had explicit policies to publish null results for over 30 years; see Broughton, 1987). Our view is that this difference has little to say about psi, but is indicative of an anti-psi bias in the mainstream publishing world, rather than a pro-psi bias in the parapsychology world.

We close with a response to Howard's (2018) assertion:

> *Given that psi research is still heavily doubted in most academic outlets, it is safe to say that psi effects are not typically noticeable to the naked eye of the careful observer. (p. 224)*

Since surveys consistently show high levels of belief in a range of paranormal phenomena among the general public, and personal experience is cited as a primary driver of that experience (Castro, Burrows, & Wooffitt, 2014; Pechey, & Halligan, 2012), it seems safe to say that *in situ* some psi effects *are* typically noticeable to the naked eye. However, we concur with Howard with respect to laboratory-based evidence for psi, where psi effects are inferred from cumulative statistical deviations from chance expectation. We thus commend the author for his willingness to investigate the sensitive topic of dream-ESP, and we note his adherence to transparent research practices, as the means by which these subtle effects can be discerned and debated. Indeed, it is this transparency that has enabled us to raise concerns about Howard's inclusion/exclusion strategy, which seems to us *post hoc* and open to expectancy bias; his estimates of the number and impact of unpublished studies, which shows a lack of familiarity with the parapsychological community and its history of publishing nonsignificant outcomes; and his concerns about the smaller *ES* with larger *n* studies, which ignores other factors that covary with that parameter and offer plausible explanations for that difference. Nevertheless, we do agree with Howard that this research topic is worthy of continued investigation, that research designs still have potential for further methodological improvements, and that the evaluation of such research would benefit from a clear consensus on the most appropriate meta-analytic approach.

## References


Borenstein, M. (2005). Software for publication bias. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 193-220). West Sussex, UK: John Wiley & Sons.

Broughton, R. S. (1987). Publication policy and the Journal of Parapsychology. Journal of Parapsychology, 51(1), 21-32.

Cardeña, E. (2018, May 24). The experimental evidence for parapsychological phenomena: A review. American Psychologist, 73(5), 663-677. http://dx.doi.org/10.1037/amp0000236

Castro, M., Burrows, R., & Wooffitt, R. (2014). The paranormal is (still) normal: The sociological implications of a survey of paranormal experiences in Great Britain. Sociological Research Online, 19(3), 16. http://www.socresonline.org.uk/19/3/16.html DOI: 10.5153/sro.3355

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

Dalton, K., Utts, J., Novotny, G., Sickafoose, L., Burrone, J., & Phillips, C. (2000). Dream GESP and consensus vote: A replication. The Journal of Parapsychology, 64(3), 242.

Darlington, R. B., & Hayes, A. F. (2000). Combining independent p values: Extensions of the Stouffer and binomial methods. Psychological Methods, 5(4), 496-515.

Howard, M. C. (2018). A meta-reanalysis of dream-ESP studies: Comment on Storm et al. (2017). International Journal of Dream Research, 11(2), 224-229.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings (2nd ed). Thousand Oaks, CA: Sage.

Kanthamani, H., & Khilji, A. (1990). An experiment in ganzfeld and dreams: A confirmatory study. In Proceedings of Presented Papers: The Parapsychological Association 33rd Annual Convention (pp. 126-137).

Krippner, S., Honorton, C., & Ullman, M. (1972). A second precognitive dream study with Malcolm Bessent. Journal of the American Society for Psychical Research, 66, 269-279.

Krippner, S., Honorton, C., & Ullman, M. (1973). An experiment in dream telepathy with "The Grateful Dead." Journal of the American Society of Psychosomatic Dentistry and Medicine, 20, 9-18.

Krippner, S., Ullman, M., & Honorton, C. (1971). A precognitive dream study with a single subject. Journal of the American Society for Psychical Research, 65, 192-203.

Krippner, S., Honorton, C., Ullman, M., Masters, R. E. L., & Houston, J. (1971). A long-distance 'sensory bombardment' study of ESP in dreams. Journal of the American Society for Psychical Research, 65, 468-475.

Majd, H. A., Ghobadi, K. N., Baghban, A. A., Ahmadi, N., & Sajjadi, E. (2014). Detecting and accommodating outliers in meta-analysis for evaluating effect of albendazole on ascaris lumbricoides infection. Iranian Red Crescent Medical Journal, 16(5), 1-6.

Orr, J. M., Sackett, P. R., & Dubois, C. L. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. Personnel Psychology, 44(3), 473-486.

Pechey, R., & Halligan, P. (2012). Prevalence and correlates of anomalous experiences in a large non-clinical sample. Psychology and Psychotherapy: Theory, Research and Practice. 85(2), 150-62. doi: 10.1111/j.2044-8341.2011.02024.x. Epub 2011 Jun 16.

Rhine, J. B. (1954). The reach of the mind. Harmondsworth, UK: Pelican/Penguin. (Original work published in 1948).

Roe, C. A., Sherwood, S. J., Farrell, L., Savva, L., & Baker, I. (2007). Assessing the roles of the sender and experimenter in dream ESP research. European Journal of Parapsychology, 22, 175-192.

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

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

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702-712.

Storm, L., Sherwood, S. J., Roe, C. A., Tressoldi, P. E., Rock, A. J., & Di Risio, L. (2017). On the correspondence between dream content and target material under laboratory conditions: A meta-analysis of dream-ESP studies, 1966-2016. International Journal of Dream Research, 10(2), 120-140.

Ullman, M. (1969). Telepathy and dreams. Experimental Medicine & Surgery, 27, 19-38.

Ullman, M., & Krippner, S. (1969). A laboratory approach to the nocturnal dimension of paranormal experience: Report of a confirmatory study using the REM monitoring technique. Biological Psychiatry, 1, 259-270.

Ullman, M., Krippner, S., & Feldstein, S. (1966). Experimentally-induced telepathic dreams: Two studies using EEG-REM monitoring technique. International Journal of Parapsychology, 2, 420-437.

Ullman, M., & Krippner, S. with Vaughan, A. (1973). Dream telepathy: Experiments in nocturnal ESP. Jefferson, NC: McFarland.

Van de Castle, R. L. (1971). The study of GESP in a group setting by means of dreams. Journal of Parapsychology, 35, 312.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta analysis. Research Synthesis Methods, 1(2), 112-125.

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

Watt, C., & Valášek, M. (2015). Postscript to Watt (2014) on precognitive dreaming: Investigating anomalous cognition and psychological factors. Journal of Parapsychology, 79, 105-107.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. Psychological Methods, 8(3), 254-274.