Questions about Questionable Research Practices[1]

Lance Storm

Adelaide University

Abstract. Parapsychologists have made considerable gains over the decades. For example, they have an informative (and constantly growing) series of meta-analyses, out of which have developed ongoing debates and critiques over methodologies. Some critiques, well-intentioned though they may be, center on so-called questionable research practices (QRPs), and it has even been argued that QRPs alone may account for psi effects. In this article, I critique an article by Bierman et al. (2016) on QRPs that finds a much-reduced but still significant overall effect in a meta-analysis of ganzfeld studies by Storm et al. (2010). The series of "hypothetical" analyses undertaken by Bierman et al. are themselves argued to be questionable. Researchers are advised to be watchful of methodological oversights that misrepresent the available data and ultimately cast parapsychological research in a dim light.

Keyword: questionable research practices, QRP, meta-analysis, publication bias, fraud, psi, anomalous cognition

Highlights:

- Questionable research practices in parapsychology and meta-analyses are disputed in the light of claims made by Bierman et al. (2016).
- Statistical simulations on ganzfeld data suggest optional stopping and optional extension are unlikely.
- Focusing on ganzfeld studies, the claim of publication bias in parapsychology is undermined given the long-standing controls against it.

1 Address correspondence to: Lance Storm, Ph. D., School of Psychology, Adelaide University. Email: **lance.storm@adelaide.edu.au**

https://doi.org/10.31156/jaex.27502

Often out of curiosity, scientific researchers follow up on various leads which they derive from a complex array of ideas, approaches, and findings in any given field, but the scope for interpretation may be so wide it leaves many doors and windows open in a house of meaning that we would rather see secured from outside threats. By that I mean researchers may massage the facts, or twist vital information, or tell half the story, or adopt questionable methods, or avoid conducting analyses that might turn things around. We do not have to assume we are always dealing with deliberate deception—after all, the primary objective of most researchers is to put things right, so one should give them the benefit of the doubt and assume their hearts are in the right place. It is on that note that I review a paper by Bierman et al. (2016). The major problem with this paper is that it is very loosely guided by a paper by John et al. (2012a), who went to the effort of actually sampling a population of psychologists ("5,964 academic psychologists at major U.S. universities"; p. 525). They had some real data to evaluate, whereas the analysis by Bierman et al. consists primarily of guesstimates based on John et al.'s statistics. On this same point of getting "real" data from the population in question, Palmer (2016a) notes that Bierman et al. (2016) did not need to make guesstimates; all they had to do was contact the authors of the ganzfeld papers to clarify with them whether certain QRPs were made. It is for these reasons that the approach taken by Bierman et al. needs to be addressed.

Bierman et al. (2016) borrowed five questionable research practices (QRPs) from John et al. (2012a) and came up with another two, and entered so-called prevalence intervals (a value range indicating how prevalent a QRP might be in parapsychological research) into simulations of the ganzfeld database taken from Storm et al. (2010). The ganzfeld is a "special type of environment (or the technique for producing it) consisting of homogenous, unpatterned sensory stimulation" to the eyes and ears of the participant, who is usually in 'a state of bodily comfort'" (Thalbourne, 2003, p. 45). Traditionally, the ganzfeld is a procedure for eliciting *psi* (i.e., psychic functioning such as extra-sensory perception in the form of telepathy, clairvoyance, or precognition). In a ganzfeld experiment, a so-called agent in one room is required to "psychically communicate" usually one of four randomly selected picture targets or movie film targets to a perceiver in another room—the perceiver is in the ganzfeld condition. The ganzfeld environment involves setting up an undifferentiated visual field by viewing red light through halved translucent ping-pong balls taped over the perceiver's eyes. Additionally, an analogous auditory field is produced by listening to stereophonic white or pink hissing noise. Typically, the perceiver's mentation is recorded and accessed later in

order to facilitate target identification. At this stage of the session, the perceiver ranks from 1 to 4 four pictures (one target plus three decoys; rank 1 = hit, or a standardized *z* score can be estimated), although in some studies each trial can contain fewer or more than four choices. This condition follows the noise reduction model, considered psi conducive because it allegedly reduces irrelevant background noise, leaving mainly the psi signal.

Although *p* values are exponentially extremely low in ganzfeld meta-analyses (*p* values up to 10^{-16}) and therefore significant, Bierman et al. suggested that the ganzfeld database would be considerably less significant ($p = .003$) if these hypothetical QRPs were present (for a more sophisticated treatment but similar finding, see Bancel, 2018). The following QRPs are taken from Bierman et al. (2016, pp. 8-10):

1. *Confirmation-to-pilot* (a confirmatory experiment is prematurely halted if results are negative; trials are dumped, the protocols are adjusted, the experiment is restarted);
2. *Pilot-to-confirmation* (a 'test-run' study, data of which are included in a confirmatory experiment if there are no problems);
3. *Optional stopping* (stopping short of a specified number of participants as soon as significance is reached);
4. *Optional extension* (testing more than a pre-specified number of participants and only stopping once significance is reached);
5. *Publication bias* (not publishing studies that produce only chance or negative results);
6. *Deciding to exclude data post hoc* (removing participants who are unsuitable); and
7. *Fraud* (a.k.a. "deceptive research"; Bierman et al., 2016, p. 10).

If present in the ganzfeld studies, these seven QRPs can inflate the ganzfeld psi effect (e.g., by polluting the data with false positives), but of course evidence of QRPs has to be found. Although few would doubt that tightening the reins on parapsychological practice is a worthy idea if it identifies sources of artifacts, the problems with the rationale underpinning that intention are manifold. Palmer (2016b) recognized this fact when he said Bierman et al.'s paper was "a monumental waste of time (and journal space)" (p. 143), given that the authors ostensibly implied the whole venture was hypothetical. Palmer, however, even disputed that implication, pointing out that he (Palmer, 2016a) previously had argued that Bierman et al. had every intention of proving that the ganzfeld database was "inflated by QRPs" (Bierman et al., 2016, p. 1), so that their study was not hypothetical at all. In their own words, they aimed to demon-

strate the ganzfeld database "would be a good example of how purely random data can be distorted into the appearance of significance" (Bierman et al., 2016, p. 2). I will consider each of the above-listed seven QRPs and question the rationale adopted by Bierman et al. (2016).

Confirmation-to-Pilot and Pilot-to-Confirmation

Confirmation-to-pilot (CtoP) starts as a confirmatory study, whereas Pilot-to-Confirmation (PtoC) starts as a pilot study. These two QRPs suggest that trials are sometimes excluded *just because* they are misses, or sometimes included *just because* they are hits, when the decision to exclude or include should only hinge on technical issues to do with experimental design protocols. Bierman et al. (2016) claim that excluded data contribute to the file-drawer problem (i.e., publication bias), whereas the inclusion of questionable data (showing a promising hit rate) is inflationary. Although the latter is quite correct for both CtoP and PtoC, one of the reasons for halting an experiment is that protocol adjustments may be needed when there is a good reason to dump trials even if there seem to be (or are) excessive misses—under those conditions, those data should not be part of the file drawer. Bierman et al. may have allowed for that eventuality, but it is not clear in their paper.

Another of their problems is they (allegedly) borrowed for their simulation a CtoP prevalence estimate from John et al. (2012a), who never featured CtoP—the figure of "40.7%" (see Bierman et al., 2016, Table 1, p. 8) seems to have come from Table S2 (John et al., 2012b, p. 15) which is for optional stopping. The actual wording is "Stopping collecting data earlier than planned because one found the result that one had been looking for" (John et al., 2012b, p. 15). And there is no prevalence figure for PtoC; only prevalence intervals that seem to be stabs in the dark.

Optional Stopping and Optional Extension

Optional stopping and optional extension can push a study into significance; all the experimenter has to do is designate a certain number of participants to test, but then keep an eye on effect size and stop short when, or extend testing until, significance is reached. Palmer (2016b) keenly observed that any study that had committed optional stopping or optional extension could be identified from *N* (i.e., the number of trial counts)—specifically, "nonround *N*s" (p. 143). Of course, a non-round count does not automatically indicate a QRP; Mainstream researchers commonly report sam-

ples with nonround *N*s because of genuine oversampling, no shows, and so on. The two-fold hypothesis of optional stopping and optional extension following Palmer's assumption has not been conducted as an exercise, but I present one here.

Because researchers rarely state that they stop short, or extend testing, one has to be ruthless and exclude any study that seems to have an unusual number of trials, which means the innocent suffer with the presumed guilty. Nevertheless, on that basis, I reduced the same 102-study database used by Storm et al. (2010) in their meta-analysis to 52 studies (these 52 had *N*s of 10, or 20, or 30, and so on, up to the largest *N* of 120). Multiples of ten are as round as one can get, so I have been extremely conservative (the results of the original study, and the new findings for the reduced rounded *N* database, can be found in the Appendix).

Even for a database almost half its original size, little has changed statistically for the crucial values. Also, about 1/4 of the studies in both databases are independently significant (about the same for the original database). These results tell us all we need to know about the efficacy of the ganzfeld treatment. So, *as an exercise*, if we ruthlessly adjust for optional stopping and optional extension, on the hypothetical assumption that these options were present as QRPs in some ganzfeld studies, we end up with a purified database with consequent effects that are even more unquestionable than before (in the Appendix, I also reduced Bierman et al.'s "S1 Dataset" of 79 studies in the same conservative way—the results are comparable).

Publication Bias

Parapsychology journals (unlike mainstream psychology journals) go to great efforts to publish studies with nonsignificant results—they tend *not* to end up in the "file-drawer." That is because the Parapsychological Association (PA) Council adopted a policy in 1975 of opposing the *exclusive* publication of studies with *only positive outcomes*. Thereby, "negative findings have been routinely reported at the association's meetings and in its affiliated publications" since that date (Bem & Honorton, 1994, p. 6; see also Honorton, 1985, p. 66). Bierman et al. (2016) actually mention this publication policy, and they state that "a non-significant outcome is not a danger to the career of the parapsychologist" (p. 8). However, they argue that parapsychologists "may be driven by a non-materialist or spiritual worldview that they try to defend" (p. 8), implying that the policy would be ignored. Palmer (2016a) makes the same point about mainstream scientists (including "psi-skeptics") who defend "theories they identify with" (p. 7). Therefore, in order to be "cautious," Bierman et al. "assume that the prev-

alence values of the use of QRPs in parapsychological research *are similar to those measured in experimental psychology* [emphasis added]" (p. 8). Bierman et al. surely realize that experimental psychologists also have worldviews to defend, but the same experimental psychologists do not have a publication policy like that of the PA. Thus, there is no level playing field. In fact, it may be argued that all the QRP prevalences and prevalence intervals used in Bierman et al.'s model are inflated. Apart from this major problem, Bierman et al. (2016) state that "JLP estimated the prevalence of [optional extension and optional stopping] to be 58% … and 22%", respectively (p. 7), but these are not prevalence figures—they are "Self-admission rate[s]" from John et al.'s (2012a) Table 1 (p. 521). Also, Table 1 in Bierman et al. presents five "JLP(%)" statistics that are supposed to be prevalence estimates, but are in fact "Admission estimates" from John et al.'s (2012b) Table S3 (p. 16). These are clear errors of commission because Bierman et al. state that their simulation will use prevalence intervals based on JLP's "*prevalence figure for each specific QRP* [emphasis added]" (p. 7). Bierman responded to these discrepancies by saying: "I think we took the admission rates as prevalence figures (as a lower estimate)" (D. J. Bierman, personal communication, December 6, 2017).

It is possible to test the likelihood of publication bias using various statistical tests. Storm et al. (2010) chose two methods to demonstrate whether publication bias was likely. The first test, using Rosenthal's (1979) so-called Fail Safe File Drawer (FSFD) formula, has been used by a number of parapsychologists since the 1980s, so Storm et al. used it for comparative purposes, whereas the other test from Darlington and Hayes (2000) is more conservative. Focusing only on the result using Rosenthal's formula, Bierman et al. (2016) observed:

Storm et al calculated that 2,414 unpublished studies were required to eliminate the overall [significant] results of the GF [ganzfeld] database. Storm et al argue that this number, given the limited resources of the field and the acceptance of publishing negative findings, is unreasonably large. (p. 9)

Bierman et al. (2016) then point out that researchers who calculate these numbers "generally assumed incorrectly that the decision not to publish is unbiased," and they also argue that the same researchers err in thinking "that the unpublished studies have zero effect size" (p. 9). It *is* true that Rosenthal's (1979) method is misleading "because it treats the inherently biased file drawer as unbiased" (Scargle, 1999, p. 22), and falsely assumes an overall effect size of zero. That is why Storm et al. (2010) also presented a more conservative estimate of up to 384 studies that *does* assume bias;

not only can these unpublished studies have a zero effect size, they can have negative values (p. 477). Bierman et al. focused on the least conservative result, thus making Storm et al. look like they were unjustifiably downplaying publication bias to the extreme, having based it on a figure so inflated that it made the ganzfeld database look too good to be true. Any reader not familiar with the Storm et al. meta-analysis would assume that Bierman et al. discovered a critical flaw in the meta-analysis.

Decision to Exclude Data Post Hoc

An experimenter's decision to exclude data *post hoc* in the ganzfeld context means unsuitable participants are removed (removal could be based on the fact that the participant fell asleep, or complained, or was late, or vexed, or uncooperative, etc.). However, if a participant produces a result that "conforms to the desired hypothesis [i.e., is a hit], the experimenter may be non-consciously less inclined to remove the subject [participant]" (Bierman et al., 2016, p. 10). Bierman et al. argue that this practice is tantamount to "fraud" (p. 10) if removal of "subjects with misses" gets *too high* (nominally over 5%)! Irrespective of whether it is fair to simulate this QRP in their model, setting an arbitrary frequency of 5% to demarcate fraud from legitimate practice is itself questionable.

Fraud

In regard to fraud, I suggest Bierman et al. (2016) have little to go on, so they rely on John et al. (2012a, 2012b) to inform them. They refer to John et al.'s admission figure of about 4.4% (actually it should be 4.34%—John et al., 2012b, p. 16), and adopt an apparent guesstimate for parapsychology of 3.4% which, it turns out, is based on one case (i.e., "one senior researcher in the 80 studies post-1985 database [who] might be guilty of deception" because he or she "had been implicated in errors in the randomization procedure"; p. 10). This one case is from a database of 29 principal investigators (1/29, or 3.4%), which yields an inflated fraud rate because they could have taken the "two studies" by the "senior researcher" out of the "80 studies" (thus, 2/80, or 2.5%).

What is questionable is (i) how "errors in randomization procedure" implies fraud; (ii) why only the principal investigators represent the "persons of interest" and not all the researchers (e.g., research assistants); (iii) why 3.4% is considered a fair and valid indicator of parapsychological fraud merely because it is "close" to John et al.'s 4.34%; and (iv) why the equally valid 2.5% is not preferred.

Conclusion

QRPs basically translate as many different ways to commit fraud when the QRP is deliberate—otherwise, one could assume they only indicate carelessness on the part of some researchers. However, the practice of assessing QRPs is questionable in itself when coupled with various oversights, errors of commission and omission, misrepresentations, and so on. It must be said that some researchers, in obeisance to a higher cause, interpret conservatism a little differently than the rest of us. When things like this happen, it is hard not to see an agenda that ignores the objective aims of a science that should tell all; an aim that should also let the pieces fall where they may. John et al. (2012a) are correct in saying "QRPs can waste researchers' time and stall scientific progress …" (p. 531), but we do have to be mindful, when checking for QRPs, that we do not punish the innocent with the guilty.

Declaration of interests: The author declares that there is no conflict of interest.

References

Bancel, P. A. (2018). Simulating questionable research practices. *Proceedings of the presented papers of the Parapsychological Association 61st annual conference* (pp. 105-115). Parapsychological Association.

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

Bierman, D. J., Spottiswoode, S. J. P., & Bijl, A. (2016). Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology, *PLoS ONE*, *11*(5), 1–18.

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

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

John, L. K., Loewenstein, G., & Prelec, D. (2012a). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

John, L. K., Loewenstein, G., & Prelec, D. (2012b). Measuring the prevalence of questionable research practices with incentives for truth telling (supplemental materials (pp. 1–17). https://journals.sagepub.com/page/pss/suppl/data-supplements

Palmer, J. (2016a). Hansel's ghost: Resurrection of the experimenter fraud hypothesis in parapsychology [Editorial]. *Journal of Parapsychology*, *80*, 5–16.

Palmer, J. (2016b). Statistical issues in parapsychology: Hypothesis testing—plus an addendum on Bierman et al. (2016) [Editorial]. *Journal of Parapsychology*, *80*, 141–143.

Rosenthal, R. (1979). The 'File Drawer Problem' and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Scargle, J. D. (1999, September 17). Publication bias (The "File-Drawer Problem") in scientific inference. https://arxiv.org/pdf/physics/9909033

Storm, L. (2025). *A new approach to psi: Re-envisioning paranormal experience as meaningful coincidence.* Routledge.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analyses of free-response studies 1992-2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, *136*(4), 471–485. doi:10.1037/a0019457

Thalbourne, M. A. (2003). *A glossary of terms used in parapsychology*. Puente.

Appendix
Optional Stopping and Optional Extension—Rounding Out *Ns*

Original Database (Storm, Tressoldi, & Di Risio, 2010)

The homogeneous database consists of 102 ganzfeld studies: mean $z = 0.81$ ($SD = 1.23$; range: -2.30 to 4.32), mean $ES = 0.135$ ($SD = 0.20$; range: -0.44 to 0.65), and Stouffer $Z = 8.13$ ($p < 10^{-16}$). *CI*s (95%) are as follows: *z* scores, [0.56, 1.07]; *ES* values, [0.10, 0.17]. Note that neither of these includes mean chance expectation (MCE)—that is, zero—which means there is no statistical likelihood that the results fall within chance parameters. Of the 102 studies, 74 (72.5%) had positive *z* scores. Twenty-seven (26.5%) of the 102 studies are independently significant ($\alpha \leq .05$).

Rounded Database ("nonround Ns" removed)

The database consists of 52 studies: mean $z = 0.71$ ($SD = 1.21$; range: -1.97 to 3.41), mean $ES = 0.130$ ($SD = 0.22$; range: -0.44 to 0.65), and Stouffer $Z = 5.08$ ($p = 1.89 \times 10^{-7}$). *CI*s (95%) are as follows: *z* scores, [0.37, 1.04]; *ES* values, [0.07, 0.19]. Note that neither of these includes MCE—that is, zero. Of the 52 studies, 34 (65.4%) had positive *z* scores. Thirteen (25.0%) of the 52 studies are independently significant ($\alpha \leq .05$).

The difference between the two mean *z* values is not significant, using the $Z_{diff.}$ (*Z*-difference) formula, $Z_{diff.} = [Z_1 - Z_2]/\sqrt{2} = [0.81 - 0.71]/\sqrt{2} = 0.07$ ($p = .472$).

In a second unrelated analysis, it is worth mentioning that Bierman, Spottiswoode,

and Bijl (2016) provide a link to their own set of 79 ganzfeld studies which they label the "S1 Dataset" (p. 4). These 79 are all the post-1985 studies they deemed suitable for analysis, having excluded the questionable studies that had "procedural weaknesses" (p. 3). Bierman et al. produced a Cohen's effect size $d = 0.14$. If we take out all nonround *N*s from their database, the dataset reduces to 37 studies, with corresponding $d = 0.12$. Again, we see little change. There is little evidence of optional stopping or optional extension.