Journal of Parapsychology, 80(1), 5–16

EDITORIAL

HANSEL'S GHOST: RESURRECTION OF THE EXPERIMENTER FRAUD HYPOTHESIS IN PARAPSYCHOLOGY

By John Palmer

In his review of *Parapsychology: A Handbook for the 21st Century* in this issue of the *JP*, J. E. Kennedy (2016) comments favorably on a chapter by Douglas Stokes, a major theme of which is that a significant number of psi experiments were likely successful because of experimenter fraud. I had a far different reaction to Stokes' chapter, so much so that I decided to devote this editorial to the topic. It has turned out to be quite lengthy for an editorial, much longer than I thought it would be when I started writing it. To justify this loquacity, I cite the fact that it has been several issues since I wrote an editorial for the *JP*, so I'm just catching up.

Hansel (1966, 1980)

I was both surprised and troubled by Stokes' chapter. I was surprised because, as regular *JP* readers are aware, Stokes is a frequent critic of psi research, but this is the first time I had seen him make a push for the experimenter fraud hypothesis. Indeed, the hypothesis has been out of fashion at least since the early 1980s. Starting in the 1960s, fraud by either gifted psi subjects or the experimenters who tested them was the major line of attack by critics of parapsychology. This was due almost entirely to the psychologist (and magician) C. E. M. Hansel, who in effect (more on these two words later) accused one subject and two experimenters of using "tricks" to falsify the data in three ESP card guessing experiments back in the 1930s and 1940s (Hansel, 1966; 1980). At the time his first book came out, I was transitioning from undergraduate to graduate school and spent two summers at J. B. Rhine's parapsychology laboratory in North Carolina, where I became familiar with Hansel's attacks, and I was outraged by them. I was brought up to believe that a person is innocent until proven guilty, and that you don't accuse someone of a crime (which is what experimenter fraud is, whether or not one can be thrown in jail for it) without evidence that they did commit it, not just that they could have. To this day I have no idea why, but I was later given an opportunity to express my anger in a review of Hansel's second book for *Contemporary Psychology*, which at the time was the book review journal of the American Psychological Association (Palmer, 1981). Of greater importance (to me at least), my anger at Hansel was one of the reasons I decided to burn my bridges in mainstream academia and become a parapsychologist.

A noteworthy feature of Hansel's polemic—which in my view makes his conduct more despicable, not less—is that he does not explicitly accuse his targets of fraud. He claims only that fraud was possible, from which he further concludes that the experiments do not provide conclusive proof of psi. This is an extraordinarily weak claim, especially for someone like me (and I strongly suspect most scientists) who recognize that evidence is a matter of degree. (It reminds me of the claims made by the tobacco companies that smoking has not been conclusively shown to cause lung cancer.) So what was all the fuss about? The "trick" was to word the text in such a way that the qualifying remarks did not register with all but the most careful reader. I'd bet that if you removed the qualifying sentences from the text and asked readers what Hansel's conclusion was about a particular experiment, any reader would answer quite confidently that Hansel was claiming that the significant results were due to fraud. That clearly was the premise of both sides in the exchanges between Hansel and the defenders of the research in the journals, and Hansel never wrote anything to correct the "misperception." However, the fact remains that all Hansel can legitimately

be accused of is *insinuation* of fraud. Why did he do it this way? My guess is to avoid a possible lawsuit for defamation or libel.

In my opinion what ended the Hansel era of criticism was an article by psychologist Ray Hyman in *Skeptical Inquirer*, the main outlet for external psi-skeptics during that period. Hyman by this time had replaced Hansel as the leading critic of experimental parapsychology. Hansel had just come out with his second (and last) book on parapsychology (Hansel, 1980), which was essentially an update of Hansel (1966). The main new feature was an application of another hypothetical fraud scenario to physicist Helmut Schmidt, a prominent parapsychologist and the father of modern random number generator (RNG) research. Importantly, and unlike Hansel's previous targets, Schmidt was currently active in the field. That was too much for Hyman, who excoriated his fellow psi-skeptic for his use of experimenter fraud allegations: "The parapsychologists, of course, see Hansel's position for what it is—a dogmatism that is immune to falsification" (Hyman, 1981/1989, p. 294). Since then, what both sides generally have thought is needed to make the case for psi has shifted from Hansel's "fraud-proof" crucial experiments (FPCEs) with gifted participants to the collective significance of multiple experiments, mostly testing unselected participants, judged for methodological adequacy by the standards applied to comparable psychology experiments. Hyman has contributed to this kind of criticism himself, including in the paper I quoted above. This collective significance is determined by meta-analysis, which also demonstrates (badly, in my view) the parallel criterion of replicability. In short, the benchmark has changed from FPCEs to meta-analyses. Unfortunately, there is reason for concern that the Dark Ages are returning.

The Ghost(s)

Turning now to the resurrection, I focus on Stokes' chapter and a companion piece by Bierman, Spottiswoode, and Bijl (BSB; 2016), which provides a more technically sophisticated version of the kind of analysis around which Stokes builds his case for experimenter fraud. One notable difference between the first and second incarnations is that Hansel was a parapsychology outsider, whereas Stokes, Dick Bierman, and James Spottiswoode are all parapsychology insiders; I know nothing about Aron Bijl. Stokes' embrace of the experimenter fraud hypothesis is perhaps understandable, as he, along with Kennedy and a few others, were instrumental in uncovering the faking of data by Walter Levy at Rhine's laboratory (Rhine, 1974), one of only two adequately documented cases of this most serious type of fraud in the history of experimental parapsychology. I consider them heroes for detecting and verifying the fraud, bringing it to Rhine's attention, and ending Levy's career as a parapsychologist.

One thing I like about both the Stokes and BSB papers compared to Hansel's writings is that we are spared the latter's elaborate cheating scenarios. Second, in the Stokes paper no potentially guilty researchers are referred to by name. The situation is more complicated re the BSB paper, a matter I address in detail later. What troubles me about both papers is that they legitimate Hansel's practice of insinuating or accusing experimenters of fraud without evidence of fraud. It's a relatively small further step to start naming names.

Stokes (2015)

Stokes offers up what might be called a "thought experiment" in which he arbitrarily divides the results of 100 psi experiments into four produced by fraud (all significant at just the .01 level), 80 produced by investigators guilty of publication bias, that is, who fail to publish their nonsignificant experiments (all significant at just .05), and 16 single psi experiments with nonsignificant results (median psi score of 0). The mean expected z score of the 100 studies is ca. 0.757, which happens to be in the neighborhood of a corresponding value of 0.570 from a meta-analysis of 72 forced-choice ESP studies (Storm, Tressoldi, & Di Risio, 2012), which we are assured is typical of the values from numerous other meta-analyses that could have been chosen. We are apparently supposed to see this as evidence that the significance of the entire experimental parapsychology database can be attributed to the processes in the model, and fraud is clearly meant to be a necessary component.

In order to give Stokes' thought experiment more relevance to the real world, I will assume that the numbers of experiments in each category were at least inspired by the results of an anonymous survey emailed to 5,964 academic psychologists at major US universities by John, Lowenstein, and Prelec (2012). Based on only a 36% return rate, they estimated "that approximately 10% [precisely, 9%] of research psychologists have introduced false data into the scientific record [at least once] and that a majority … have engaged in questionable [research] practices [QRPs]…" (Stokes, 2015, p. 45). The estimated *prevalence* of each QRP was the geometric mean of (a) the percentage of respondents who admitted they had engaged in the QRP, (b) the average of the percentage of other psychologists they thought had engaged in the QRP, and (c) the average of the percentage of these other psychologists they thought would admit to the QRP. We are offered no explanation of why we should consider the respondents' last two estimates anything other than wild guesses, or to put it another way, why we should take John et al.'s prevalence estimates seriously.

In any event, Stokes' 4% of *experiments* is a reasonable equivalent to John et al.'s 9% of *experimenters*, because it is likely that the fraudulent experimenters published at least as many successful nonfraudulent experiments as fraudulent ones. To be fair, it is this and related studies, rather than his own analysis, that Stokes says converted him to psi-skeptic. I will go further and suggest that it is the John et al. study, combined with the difficulty psi-skeptics have had in explaining away by more conventional means the successful psi experiments published in a prominent mainstream psychology journal by the prominent social psychologist Daryl Bem (2011), that is responsible for the recent resurrection of the experimenter fraud hypothesis in parapsychology.

So we now have argument by analogy, which immediately raises the question, how good is the analogy? Chris Roe, current president of the Parapsychological Association, shares my alarm about the resurrection of the experimenter fraud hypothesis enough to publish an article on the subject in the Association's magazine (Roe, 2016). He argues that on a priori grounds one would expect fraud to be much less likely in parapsychology than in mainstream fields. Compared to the mainstream, successful research in parapsychology brings much less tangible reward in terms of salary and career advancement; in fact, in the U.S. at least, your chances of getting a job at a reputable university with parapsychology on your résumé are close to nil. Although you will have a good reputation among your fellow parapsychologists, it will be just the opposite where it counts, among mainstream scientists. Your research will be more carefully scrutinized, and if you are caught cheating, the psi-skeptics will make sure the whole world knows about it. BSB acknowledge the scrutiny point, but they single out parapsychologists as being driven to defend "a non-materialist or spiritualist worldview" (p .14). Although most mainstream scientists aren't driven to defend the materialist worldview (in their mind there's no need to), many are driven to defend lower-level theories they identify with. One group of mainstream scientists that does see a need to defend their worldview is the psi-skeptics, and they do so at least as passionately as psi-proponents. In short, by this criterion it's at best a tie.

However, the most important point to make is that evidential claims based on analogy are inherently much weaker than claims based on empirical evidence. Indeed, Stokes' thought experiment is about as far from a "smoking gun" as one can get. Fraud is a serious charge, and if a prosecuting attorney presented anything close to Stokes' analysis as evidence of guilt it would be laughed out of court. But even if we dignify Stokes' thought experiment as a scientific test of a scientific hypothesis, it fails on those terms. For a finding to support a theory, it must not only follow from that theory, but it also must *not* follow from a straightforward application of the competing theory (i.e., without the addition of gratuitous corollaries). In this case, it is obvious that Stokes' result could be explained just as easily by genuine psi as by fraud and publication bias, so its evidential value is nil—unless one follows Hansel (1966) by arguing that any other hypothesis should be chosen over psi on a priori grounds, the logic that prompted Hyman (1981/1989) to declare Hansel's position unfalsifiable in the quote above.

Bierman et al. (2016)

BSB analyzed six methodological flaws, which, following John et al. (2012), they call "questionable

research practices." Two of these, "fraud" ("deception") and "publication bias," are the ones Stokes had in his model. The four new ones all concern misclassification of participants as in or not in the formal sample: "confirmation to pilot," "pilot to confirmation," "optional stopping," "optimal extension," and "biased removal of Ss [participants]." The criterion sample was a combination of 102 ganzfeld studies from two meta-analyses (Storm & Ertel, 2001; Storm, Tressoldi, and Di Risio, 2010), cut to 78 by removing studies published prior to 1985. Whereas Stokes started with a single model and tried to find a criterion sample that fit it, BSB started with a single criterion sample and tried to find the model that fit it best. The goals of the analyses also were different. All Stokes asked of his model was to predict the average effect size in the criterion sample. BSB had the more ambitious goal of seeing if they could explain away all the statistical significance of the combined studies in their sample as due to the QRPs. They didn't quite get there, but they came close.

The BSB analysis is an improvement over the Stokes analysis in that the inductions from John et al. (2012) are explicit and reasonably well-defined. But in contrast to those of Stokes, BSB's conclusions about the proportions of QRP studies in their models were also influenced by their judgments of whether QRPs were committed in the individual experiments in the database. If these were straightforward evaluations, I would have endorsed this procedure as an unqualified improvement over Stokes'; but instead they give us another layer of inferences—inferences that are more invalid (and more unethical) than the inferences from John et al.

Are the QRPs fraudulent? The only QRP besides "fraud" that BSB explicitly label as fraudulent is biased removal of participants, but only in some cases: "Removing a larger percentage than 5% of subjects with misses … will make the post hoc arguments for removal that an experimenter has to come up with increasingly artificial and will basically turn this practice into fraud …" (p. 20).[1] It is unclear whether only the "fraudulent" cases were included in the model.

As for the other QRPs, it is important to note that what makes a QRP fraudulent is not its nature but its intent. Take for example an (incredibly bad) experimenter who has a (fake) psychic guess a sequence of cards exposed one by one on a table he is seated at, while wearing a blindfold. The psychic gets a high score. If the experimenter is totally naïve about magic tricks and thought that the psychic couldn't see the cards (e.g., by looking through the small gaps along the sides of the nose) we would classify the artifact as "sensory cues." If, on the other hand, the experimenter knew the psychic could see "through" the blindfold, went ahead anyway, and failed to mention the artifact in the report, we would call it experimenter "fraud."

Thus, whether we can say that BSB's claim that these other QRPs are fraudulent depends on whether they are claiming the experimenter knew it was wrong. I will argue that for at least one QRP, optional stopping, BSB assume that at least the vast majority of the investigators in their sample knew it is wrong, based on the statement that "…students are generally taught that this particular flexibility in experimental practice gives misleadingly inflated scores" (p. 12). Second, it is important to note that optional stopping can only bias the data if the investigator is keeping track of the results and stops because he/she knows that stopping at this point will produce a favorable outcome. That this was assumed to be the case is evidenced by the following statement, which was cited in reference to another QRP but in fact applies to all the misclassification QRPs: "… we choose in our model for this QRP that the researcher starts contemplating early stopping and checking whether the *p*-value is smaller than 0.05 at trial 10" (p. 32). This implies intent, hence fraud. But what if the investigator is not keeping track of the results? I purposely keep myself blinded to how well an experiment is going because I am afraid knowing this might affect how I interact with future participants. The point is that my stopping a study at a certain point absent knowledge of previous scores has equal chances of helping and hurting the final outcome. In fact, one of the reasons I keep myself blinded is so I can stop a study prematurely if I want to or need to for some legitimate reason. BSB recognize this: "If subject removal happens blindly before inspection of the data, then the practice will not introduce a bias" (p. 19). The problem is with the immediately following sentence, which appears to be the basis for

[1] The online article has no page numbers. This and subsequent page numbers are from the preprint Bierman sent me, which is 33 pages up to the references. The page numbers should give the reader a rough idea where to find the quotes in the online article. *PLoS ONE* is open access.

inferring that all the optional stopping cases in the database are genuine QRPs: "However in GF research experimenters are generally not blind to the outcome of a session" (p. 19).

Are the "QRPs" QRPs? What disturbs me most about the BSB paper are the unjustified inferences that a QRP was committed at all. If BSB were interested in whether the "QRPs" were *really* QRPs they would have (a) carefully read the Method sections of the reports, and (b) if it was still unclear whether the procedure was defensible, contact the author. Many of the authors in the list of studies are currently active and accessible. Finally, (c) they would have removed the QRP tag from the experiment if an adequate justification was presented. There is nothing in the article to suggest that BSB did either (b) or (c); if they did, they certainly would have mentioned it. Of course, following this path would have reduced the percentage of studies identified as QRP studies, and adjusting the parameters of their model accordingly would have reduced the likelihood of explaining away all the significance in the ganzfeld database.

Optional stopping. I was most troubled by BSB's treatment of two of the QRPs. The first is optional stopping, the presence of which they infer from "publication of a round number of trials" (p. 13). This is doubly egregious: not only can't you infer that the stopping was intentional (as discussed above); you can't even infer that it was optional (i.e., not specified in advance). There are many "innocent" reasons for a nonround number of participants. Although this round number criterion was mentioned in a paragraph devoted to optional stopping, it applies equally to the other misclassification QRPs, so I assume it was used for these as well. My use of the list of studies was complicated by two other deficiencies in the report: (a) "Round number" is never adequately defined; some guidance is given by the statement "N = 10, 20, 30 etc. with frequencies around 10" (p.13), but it's still not clear if all numbers not ending in 0 are "nonround"; I think most people would consider 25 to be round in the sense BSB seem to mean. (b) The numbers of studies in each QRP category are not listed for any of the models.

Anyway, I went to the list to try to identify the misclassification studies based on whether the sample size looked odd to me, and to my astonishment I found one that had me listed as an author (Kanthamani & Palmer, 1993)![2] Although Kanthamani was listed as first author, the study was primarily mine, and I was listed as first author in the earlier version presented at a Parapsychological Association convention (Palmer & Kanthamani, 1990). I was the one who made all the important decisions pertaining to the analyses. I was puzzled myself why N was 22, as that is not a number I would ordinarily prespecify, so I raced to grab a copy of the report, where I found the following paragraph in the "Subjects" section under "Method":

> We tested more than 20 Ps [percipients – they brought their own senders] because two pairs were able to complete both sessions after we had assumed they could not, and we thought it best to fulfill our commitment to test them. No other factors influenced our decision to test 22 Ps. None of the mentation reports were judged [outside judges only] or ESP scores computed until all subjects had been tested. (Kanthamani & Palmer, 1993, p. 244)

In retrospect it would have been better to use "test and analyze" in the penultimate sentence but I think it is clear from the context, especially the last sentence, that this is what was meant. More importantly, the paragraph makes it clear that I was aware that optional extension could be a QRP, and I took steps to do the procedure in a way that would not bias the results. I can't be 100% certain that my study was classified as a QRP study, but I consider it extremely likely that it was. If so, I see two possibilities, both of which are damning to BSB: (a) They didn't care if the procedure was justified so they didn't read past the abstract, or (b) they read the passage quoted above but still classified the study as a QRP, which is tantamount to accusing me of lying. All this was inferred from an irregular sample size!

To see if I could get an example of an investigator specifying a nonround number of participants at the outset, I chose an experiment by Roe, as I was planning to send him a draft of the editorial and could at the same time ask him for an explanation of the nonround sample size if necessary. The example I chose was listed as "Roe…Sherwood…and Holt … 2004" with an N of 17. When I looked at the list a bit more closely I discovered a second study with the same designation except that N was 23. I was able to locate the

[2] Technically, this study should not even have been in the database. When Storm et al. (2010) homogenized their database, reducing it from 108 studies to the 102 that BSB drew from, this was one of the studies they removed.

report (Roe, Sherwood, & Holt, 2004), where I found that the total *N* was in fact a nice round 40. What BSB had done was to treat the sender condition ($N = 23$) and the no-sender condition ($N = 17$) as separate experiments. Nowhere in the report is this apparent departure from standard procedure mentioned (as it should have been), nor is there any mention of such in the report of the source meta-analysis (Storm et al., 2010). The statistical test for the entire sample was given in the report of the experiment (Roe et al., 2004), so it was not necessary for BSB to split the sample. What BSB got from the split was two more studies to place in the QRP bin, whereas they would have gotten none had they treated the study the same as all the others in the database; I could find no other examples of a split in the list of studies. Actually, this splitting should have been applied to all studies in which the sample was divided into two subsamples. I very much doubt Roe et al. (2004) was the only one. If most of these other studies had round subsample *N*s, BSB introduced a major source of bias into their models.

The question, then, becomes why the condition *N*s weren't split 20-20. Roe explained that immediately before the session the computer made an open-deck random decision (which does not guarantee an even split) as to whether to assign the pair to the sender or no-sender condition. The purpose was to "ensure that no-one but the nominal sender themselves was aware of the condition type until all data were collected and recorded" (C. Roe, personal communication, June 12, 2016). He also noted that computer assignment of the condition designations was mentioned in the report. Finally, this exercise suggested to me that BSB were in fact perusing the Participants sections of the reports, which increases the likelihood that the "lying" interpretation of BSB's conclusion re the Kanthamani and Palmer (1993) experiment is the correct one (see above).

Fraud. The second QRP treatment that concerned me, unsurprisingly, is the one labeled "fraud."[3] Here is what BSB say they did:

> We decided on the basis of the arguments given above [presumably the proportion of fraudulent researchers in the John et al. (2012) survey and the documented cases of fraud in parapsychology] that *one senior researcher* in the 80 studies post-1985 database might be guilty of deception. In order to take into account the contribution of deceptive research to the database we thus removed the two studies of *one senior researcher*, the person who had been implicated in errors in the randomization procedures. These studies were quite significant with HRs [hit rates] of 35% and 41%. In the database there are 29 principal investigators, so removal of one of them (3.4%) is near identical to the prevalence of deception found by JLP (4.4%). (Bierman et al., 2016, p. 21; italics added)

It looks on the surface that they are accusing one researcher of fraud, but a careful reading of the above quote paints a more complicated (and confusing) picture. First, exactly how was the experimenter selected for this dubious honor? It appears that one criterion is that the person fit the requirement of the model (which is outrageous from the ethical standpoint). Assuming we are talking about first authors, (a) the researcher had to be someone who had at least two experiments in the database and (b) both of these experiments had to have high HRs (Why else would they cheat?) As I was able to access the list of experiments, which included their HRs, I was able to identify the target by matching the HRs in the list to those in the quote. (For obvious reasons, I will not name this person, who is still active in the field.) I noted two relevant items. First, this author had more than two studies in the database. Second, there were other researchers who also had two high effect size studies. So how were the two target studies chosen? This presumably is where the randomization comes in. It would be plausible to conclude that BSB read through the candidate reports and chose the two studies with the most questionable methodology, which in this case happened to involve randomization.

There are two details about this paragraph that raise questions about what is really going on. First, it's hard to understand why BSB removed a QRP study from the database. My understanding of the procedure was that inferences from the database would take precedence over inferences from John et al. (2016) if

[3] The transgressions subsumed under this heading and those labeled as fraud by Hansel and Stokes seem more egregious than the others, so I suggest that they be called "hard fraud" and that those such as publication bias and the misclassifications be called "soft fraud."

the two estimated percentages differed. Also, as I noted above, elsewhere it seems that BSB were going out of their way to add QRP studies to their models. At least this move does make me feel a little better about their objectivity vis-à-vis the models (but not the insinuations/allegations of fraud). This is just one more example of how poorly written this paper is with respect to communicating what was done.

Second, how many researchers are we talking about? Note in the quote that the person referred to in the first sentence and the person referred to in the second (and third) sentences are both labeled "one senior researcher." However, different things are being said about them, so they are not necessarily the same person. You would expect the authors to recognize the juxtaposition and to clarify things by, the second time, using "this researcher" if the researchers are the same and "a different researcher" if they are different. So either this is just another example of inept writing or the authors intended to create ambiguity in the mind of a reader careful enough to notice the problem.

Why might BSB want to obfuscate? At the beginning of this section, I wrote "*on the surface* they are accusing one researcher of fraud." The qualifier is necessary, because the authors are very careful not to go that far. Note that they say senior researcher 1 "*might* be guilty of deception" and the transgressions of senior researcher 2 were "*errors* of randomization" even though they were *classified* as fraud. The authors are (apparently) covered regardless of whether there are two researchers or one, and the ambiguity about the number just adds another layer of confusion about who is being "accused" of what. So all I can claim is that the authors are *insinuating* fraud, following the lead of Hansel. I expressed the opinion in that earlier discussion that an insinuation is at least as reprehensible as an accusation, and BSB have even less grounds for it than Hansel did.

There were some curious goings-on that could possibly be interpreted as suggesting that someone was concerned about a lawsuit re the BSB article. We must assume that at the time the paper was accepted for publication nobody thought there was a problem because nowhere in the body of the paper are any proper names associated with the QRPs. But what if late in the game it dawned on someone, perhaps as the result of my inquiries, that there might be a problem if those names could be deduced from information available elsewhere, namely the list of studies in the supplementary information section? Of course, this is what I consulted to identify the "culprits" in my two examples. These remarks supply the context for what I present below. From here on I will stick to the facts and let readers draw or not draw their own conclusions from them. Perhaps Bierman will clarify the situation in his anticipated reply to this editorial in the next *JP*.

In late April I asked Bierman to e-mail me a copy of the full report of the analysis he had submitted for the Parapsychological Association convention. Instead, he sent me a copy of the paper that had been accepted for publication in the online open access journal *PLoS ONE*. This was fine, except for the fact that missing were the figures and three supplemental files that were identified but not included. One of the latter, labeled "database," appeared to be the list I was particularly interested in. So I wrote him back asking him to send me the figures and supplements. He refused, without explanation, telling me instead that the article should soon appear in *PLoS ONE* and I should access the materials there (D. Bierman, personal communication, April 29, 2016). This struck me as odd, as it would seem to be a simple matter to just send the files as e-mail attachments. On or around May 6, I went to the *PLoS ONE* site and found that the paper had been uploaded just a couple days previously. However, when I clicked on the icons for the supplements I got an error message telling me they were not available. I assumed this was just a technical glitch due to the recency of the upload and it would soon be identified and resolved. Sure enough, when I tried a few days later I was able to download the supplements, and the database file was the list I expected. I assumed the fix was permanent so I didn't print out the database file, as I was not ready to use it. But when I tried to access the file a week or so later I got the error message again. I wrote a letter of complaint to the journal office and promptly got a very nice reply including the files as attachments. But as of July 18, 2016 they were still not accessible on the website. The other curiosity is that the list, an Excel spreadsheet, listed the ganzfeld articles only by author last names and year. There were no journal names, nor volume and page numbers. The reason you supply such a list is so that readers can find and read the articles that were included in the analyses (cf., e.g., Storm et al., 2010). Why was this information omitted? I wrote back my contact at the journal office asking if there was a limit on the size of supplement files. She replied that there was, but it left ample room for BSB to supply complete references had they so desired.

Inferences from psychology. I will close this section with one additional observation about the BSB report. In the abstract, one finds the following statement: "Restricting the parameter space to ranges observed in studies of QRP occurrence, under the untested assumption that parapsychologists use comparable QRPs, the fit to the published Ganzfeld meta-analysis with no anomalous effect was poor" (p. 2). This seems to suggest that one of their models was an indirect test of how well the hit rate in the database could be predicted from the QRP prevalences in John et al. (2012), although I could find nothing in the body of the paper that either supports or refutes this interpretation. If my interpretation is correct, the quoted outcome is evidence against Stokes' premise (granted by me) that the prevalence of QRPs should be the same in parapsychology as in psychology.

Conclusion

The core problem with both the Stokes and BSB analyses is quite simple: You can't conclude anything reliable about the existence of fraud from inference in the absence of evidence. Insinuations or allegations of fraud are a serious matter and not something to play mathematical games with, especially when there is any chance that the target persons can be identified. Fraud must be detected, not inferred. This means that it has to be determined on a case-by-case basis, by reviewers looking for possible evidence of QRPs in the reports and following up with the author, and by lab associates who happen to notice irregularities in a colleague's data (not by conducting a witch-hunt!) and following up on their observations, as Stokes and Kennedy did in the Levy affair (Rhine, 1974).

So far I have been concerned with how past fraud should and should not be determined. I turn next to how future fraud might be detected.

A Solution? The Heymans Project

Both Stokes and Bierman maintain that the way to address the experimenter fraud problem is to conduct experiments in a way that precludes it, partly by monitoring the activities of the experimenter. It is represented by an ambitious proposed series of multilab experiments to be arranged by the Heymans Anomalous Cognition Group (2016). Bierman is listed as one of three "core members" of the group, but he seems to be the de facto head of the operation (e.g., he is identified as the one to whom inquiries should be addressed). I first learned of the project when Bierman asked me to contribute to the development of a test (stage 1 of the project) to select participants that would maximize the likelihood of success for the stage 2 experiment, most likely a replication of one of Bem's (2011) "notorious" precognition experiments, conducted in multiple labs all presumably headed by psi-proponents. If this experiment succeeded (provided evidence for psi) it would be repeated to see if the results would hold up if the labs were "adversarial," which presumably means headed by psi-skeptics (stage 3).[4]

Will the Project Demonstrate Psi?

Before I explain why Bierman's protocol cannot eliminate the possibility of fraud, I want to explain why I think it unlikely to demonstrate psi even if it could eliminate fraud. I have developed a crude formula that encapsulates what I believe is required for someone to demonstrate psi: psi strength (P) = psi ability (A) x motivation to see psi demonstrated in the experiment (M), or P = AM. There are four corollaries: (a) a P score below a certain threshold level means no psi, (b) the person must be psychologically involved in the experiment, (c) psi can manifest over long distances (see, e.g., McMahan & Bates, 1954), and (d) there need not be a conscious intention to use psi to influence the results; that is, it could be *implicit psi*, which has been demonstrated, for example, by Bem's (2011) experiments. This model has led me to conclude that at least some of the psi in most psi experiments is caused by the investigators ("experimenter psi" or E-psi), because we have good reason to believe the successful ones have high scores on both A and M in my formula (Palmer & Millar, 2015).

[4] This description of the protocol is taken from Bierman's letters to me, referenced below. A somewhat more detailed parallel description can be found on the website (Heymans Anomalous Cognition Group, 2016).

I believe this model also would hold true for Bierman's experiments. I think most psi-skeptics would agree that if there were to be genuine psi in Bierman's experiments, it most likely would be E-psi (see, e.g., Kennedy, 2016). But what about the psi-skeptics? I suspect that most of us think it unlikely that a psi-skeptic would have any psi ability. At least part of the reason, I suspect, is that we think psi-skeptics invariably haven't had psi experiences, because if they had they wouldn't be skeptics. Indeed, if you ask psi-skeptics, "Have you ever had a psi experience?" the answer is invariably "no." But that could be an artifact of the way the question is typically phrased. Psi-skeptics will interpret this question as asking whether they have had an experience that was caused by psi. Of course the answer is "no": they are skeptics. This is also likely to be the case for any phrasing that does not explicitly deny this meaning. But if you asked, "Have you ever had an experience that looks like psi on the surface, whether it is really psychic or not?" The honest answer to this question may well be "yes."

At this point, I will share my "theory" of what drives the skepticism of at least some of the strong psi-skeptics of the type appropriate for a role in stage 3 of the Heymans project. Here goes. They have had in the past one or more impressive "psi" experiences that they could not easily explain away. These experiences were and are deeply troubling to them, because they conflict with their view of the world, the validity of which is very important to them. Thus, they are strongly motivated to find a conventional interpretation of their experiences. Because they have scientific skills and training, and given the difficulty in explaining their experiences away on their own terms, they seek to fulfill their objective by applying their skills to prove that psi in general does not or cannot exist. Following this line of thought, we would expect such psi-skeptics to be fans of Bayesian statistics because it allows them to prove that the null hypothesis is false (no psi).

This "theory" is pure speculation; I have no evidence for it. However, I am not aware of any evidence against it, and personally I believe on plausibility grounds that it is true in some cases. I normally don't like to "psychoanalyze" other scientists in print, even if I don't name them. My reason for the exception is to make the point that, *if we assume that psi exists*, my speculation provides a plausible basis for suspecting that the psi-skeptics in the Heymans project, who would have high psi strength scores in my formula, would unconsciously and unintentionally exert a negative E-psi effect that would offset any positive psi exerted by the psi-proponent investigators and/or participants.

Will the Project Preclude Fraud?

The controls against fraud in the Heymans project are to automate the procedure as much as possible and "to have outside skeptics controlling our measures like real time uploading of data…"(D. Bierman, personal communication, February 3, 2016). I found it noteworthy that nowhere in his description did Bierman say anything about the need to control the psi-skeptics, who could cheat by embedding malware in the program during its creation (if they had a role in that) or manipulating the data in real time while they are doing the "controlling." So I sent Bierman an e-mail asking what he planned to do to control the psi-skeptics. He took my question seriously and politely responded that he did plan the controls I asked about, although he hadn't worked out the details (D. Bierman, personal communication, May 19, 2016). I find that encouraging, but we'll have to wait to see if he follows through.

Unfortunately, even if Bierman does follow through it wouldn't solve the problem, because you would need someone to monitor the person monitoring the controller to be sure that he/she did his/her job properly, but then you would need someone to monitor this person, and so on. In other words, there is an infinite regress. This is why it is impossible to eliminate the possibility of fraud entirely in the Heymans project. In theory you could come close by following the suggestion of Stokes (2015), that is, by at each stage of the process "having multiple parties observe the experiment" (p. 46), the idea presumably being the more pairs of eyes, the more likely it is that one of them will catch the fraud. The problem is that what's important is not quantity but quality, and whether the quality is adequate is hard to verify a priori. I am reminded of the debate parapsychologists had in the 1970s about whether magicians should be present when we test psychics such as Uri Geller. Being the rigorous sort I was firmly in the "yes" camp. However, when

it looked like we might actually do this, the deflating response we got from the (magician) psi-skeptics was "magicians fool each other all the time." Good point, actually. Although for the Heymans project the relevant skills are computer skills rather than magic skills, the same argument applies.

This last alternative highlights my primary objection to the kind of solution Stokes and Bierman propose. It creates a climate of paranoia that can be highly toxic and adversely affect the conduct of research at multiple levels. No researchers are going to be comfortable conducting an experiment with people who probably think they are cheats "looking over their shoulder," even from a distance. If I were an experimenter in the Heymans project, I would be scared that the psi-skeptic observer would implant malware into the software that would make me look like a cheat, and the psi-proponent chosen to monitor the psi-skeptic wouldn't be skilled enough to detect the manipulation. If that sounds paranoid, that's the point. The experimenter's discomfort could be picked up by research participants, adversely affecting their task performance, and it could even degrade the experimenter's ability to properly perform the steps of the procedure. However, the most damaging objection is broader in scope: Who would want to become a scientist if it means working in that kind of environment? Perhaps this is why we have not seen mainstream sciences adopt such procedures.

The Real Solution

Before presenting my solution, I must address the question of why we want a solution. I think most parapsychologists would agree that our goal is to be able to report successful experiments of such quality and integrity as to convince mainstream scientists that psi is real. I agree—with one caveat. I do not believe it is necessary to convince those within that population who are strong psi-skeptics, which most parapsychologists agree is a hopeless task. Although it may not seem this way because they are so "loud," strong psi-skeptics are most likely a minority within the scientific community. If the "silent majority" of scientists with an open mind about psi's existence were to read the evidence and thus be persuaded of psi's reality, and that majority included a proportionate number of scientists of distinction, that would be sufficient to define the "corporate position" and get parapsychology the associated tangibles such as mainstream grants and faculty appointments. (The downside, ironically, is that mainstream scientists would likely rush into the field and sweep us old-timers aside.) In any case, this is why the solution need not incorporate procedures intended specifically to convince psi-skeptics.

There is no need for the fraud-proof crucial experiment (FPCE) approach to assure that the knowledge base is not contaminated by false data, because science for years has agreed on a different solution: independent replication. In accepting this solution, they implicitly reject the FPCE approach, because they recognize that there are simply too many unknowns in any single experiment to put much faith in it, even if it somehow could be shown to be fraud-proof. However, for replication to achieve the objective, the replicating experimenters must be the kind of people highly unlikely to commit fraud. This requires that they have no incentive to commit fraud, which in the case of parapsychology is best achieved by removing from consideration individuals biased either for or against the psi hypothesis. In other words, they should come from the same subpopulation of scientists whom I suggested above we need to convince that psi is real.

The idea has been kicking around for years in parapsychology that psi might be inherently unreplicable, and I heard Bierman endorse it in a workshop at the 2015 PA Convention. It is even "predicted" by a major physics-based theory in the field (Lucadou, 1995). I think there are too many examples of albeit insufficient replication in parapsychology for this to be at all likely, but if it's true we'll just have to face the fact that we will never be able to make a scientifically compelling case for psi. Trying to get around it by turning parapsychology into a paranoid science is a fool's errand doomed to fail.

I want to end this negative essay on a relatively positive note by pointing out that the Heymans project actually fits the required mold rather well, because the multilab experiments are close-to-independent replications not only of one another but also of Bem's (2011) original experiments.[5] So, here is how I suggest Bierman modify the Heymans project if he would like to maximize the chances of providing evidence for psi convincing to the "silent majority" of mainstream scientists.

[5] Given Bierman's attitude toward replication, it is odd that he would propose a replication project.

1. Abandon stage 3. *Psi-skeptics should have absolutely nothing to do with the project.*
2. Select a minimum of five independent Principal Investigators (PIs) for the stage 2 experiments. They (along with the members of their research teams) would need to meet the following criteria: (a) the PIs should be senior psychologists who are widely respected by the community of scientists (probably, but not necessarily psychologists) for their scientific accomplishments and integrity; (b) they should believe that the evidence, insofar as they are aware of it, is insufficient to justify either a positive or negative conclusion about the reality of psi (although ideally they would have genuine curiosity about whether psi is real); (c) they should have no preference for a positive or negative outcome; (d) they should have no prior involvement with anything having to do with psi or related anomalies, including publications; and (e) the PIs should have no fear of the repercussions if they obtain positive results. If they do, they can expect to be publically excoriated by the psi-skeptics. Thus, they must be secure both internally (ego strength) and externally (job security). Perhaps the greatest challenge would be to assure that a researcher claiming neutrality really is neutral. That is clearly impossible in any strong sense, but the odds are better if the person chosen has an applied rather than a theoretical orientation; if someone with a theoretical orientation is chosen, the theory should not be grounded in physicalism. I'm not suggesting it would be easy to find these PIs, but it would help immensely if they could be offered grant money.
3. Arrange for a limited number of psi-conducive experimenters, and possibly carefully selected psychics with demonstrated psi ability, to be "psychologically involved" in the experiment in ways that can't compromise the experiment's integrity. The PIs should establish (long-distance) contact with them, keep them informed about the progress of the study, and even solicit advice about aspects of the procedure consistent with the protocol, such as how to interact with participants. They should definitely know when each session is to be conducted. In turn, these psi-proponents have the responsibility to create a good impression on the PIs and not be offended if their advice is not taken.
4. Last but not least, make sure that the protocol includes rigorous controls against all nonpsi artifacts except experimenter fraud.

Finally, the replications in this adaptation are more fully independent than in the Heymans project because the E-fraud monitors, which by necessity would have to be mostly the same for each experiment, are eliminated. It is true that the psi-proponents in the adaptation would also most likely not be independent, but their effective role would be more like that of a subject than an experimenter.

In short, mission accomplished.

References

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

Bierman, D. J., Spottiswoode, J. P., & Bijl, A. (2016, May 4). Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology. *PLoS ONE*. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0153049

Hansel, C. E. M. (1966). *ESP: A scientific evaluation*. London, UK: MacGibbon & Key.

Hansel, C. E. M. (1980). *ESP and parapsychology: A critical re-evaluation*. Buffalo, NY: Prometheus.

Heymans Anomalous Cognition Group (2016). Retrieved from https://sites.google.com/a/rug.nl/heymans-anomalous-cognition-group/

Hyman, R. (1989). Further comments on Schmidt's PK experiments. In R. Hyman, *The elusive science: A scientific appraisal of psychical research* (pp. 289–295). Buffalo, NY: Prometheus. (Reprinted from *Skeptical Inquirer*, *5*(3), 34–41).

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

Kanthamani, H., & Palmer, J. (1993). A ganzfeld experiment with "subliminal sending." *Journal of Parapsychology*, *57*, 241–257.

Kennedy, J. E. (2016). [Review of the book *Parapsychology: A handbook for the 21st century*, edited by E. Cardeña, J. Palmer, & D. Marcusson-Clavertz]. *Journal of Parapsychology*, *80*, 99–106.

Lucadou, W. v. (1995). The model of pragmatic information (MPI). *European Journal of Parapsychology, 11*, 58–75.
McMahan, E. A., & Bates, K. Jr. (1954). Report of further Marchesi experiments. *Journal of Parapsychology*, *11*, 82–92.
Palmer, J. (1981). [Review of the book *ESP and parapsychology: A critical reevaluation* by C. E. M. Hansel]. *Contemporary Psychology*, *26*, 9–10.
Palmer, J., & Kanthamani, H. (1990). A ganzfeld experiment with subliminal "sending." *Proceedings of Presented Papers: The Parapsychological Association 33rd Annual Convention*, 227–242.
Palmer, J., & Millar, B. (2015). Experimenter effects in parapsychology research. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A handbook for the 21st century* (pp. 293–300). Jefferson, NC: McFarland.
Rhine, J. B. (1974). A new case of experimenter unreliability. *Journal of Parapsychology*, *38*, 215–225.
Roe, C. (2016). The problem of fraud in parapsychology. *Mindfield*, *8*(1), 8–17.
Roe, C. A., Sherwood, S. J., & Holt, N. C. (2004). Interpersonal psi: Exploring the role of the sender in ganzfeld GESP tasks. *Journal of Parapsychology*, *68*, 361–380.
Stokes, D. M. (2015). The case against psi. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A handbook for the 21st century* (pp. 42–48). Jefferson, NC: McFarland.
Storm, L., & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin*, *127*, 424–433.
Storm L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, *136*, 471–485.
Storm L., Tressoldi, P. E., & Di Risio, L. (2012). Meta-analysis of ESP studies, 1987–2010: Assessing the success of the forced-choice design in parapsychology. *Journal of Parapsychology*, *46*, 243–273.

Acknowledgements

I would like to thank Stephen Braude, Mark Leary, and Chris Roe for their comments on an earlier draft of this editorial. Any remaining deficiencies are in no way their responsibility.

Rhine Research Center
2741 Campus Walk Ave.
Building 500
Durham, NC 27705, USA
john@rhine.org