Journal of Parapsychology, 80(2), 141–143

EDITORIAL

STATISTICAL ISSUES IN PARAPSYCHOLOGY: HYPOTHESIS TESTING—PLUS AN ADDENDUM ON BIERMAN ET AL. (2016)

By John Palmer

Hypothesis Testing

In a previous editorial I described and defended my heretical views on how multiple analyses of empirical results should be addressed (Palmer, 2013). In this editorial, I express and defend equally heretical views on hypothesis testing. The issue of how hypotheses should be evaluated statistically is important for two reasons. First, confirmation of a hypothesis goes beyond confirmation of the effect itself because it supports, or at least should support, a theory or model. Second, a more lenient criterion of statistical significance is commonly applied to hypothesized effects than to other effects, which are often labeled "post hoc." I have heretical proposals regarding both of these observations.

A tagline for my first proposal is that what is important is not whether an outcome is *hypothesized* but whether it is *hypothesizeable*. There are two key circumstances where there is a mismatch between the two. The first is: *An outcome is hypothesized that should not have been hypothesized (i.e., is not hypothesizeable)*. The fact that hypothesis tests are supposed to be tests of a theory or model implies that the author has an obligation to show how the hypothesis follows from the theory and/or previous empirical results related to the theory. In fact, this is one of the major purposes of the introduction section of a research report. This prescription is expressed as follows in the Publication Manual of the American Psychological Association (2010): "In empirical studies, [explaining your approach to solving a problem] usually involves stating your hypotheses or specific questions and *describing how these are logically connected to previous data and argumentation*" (p. 28; my emphasis). The word "argumentation" leads me to point out that the theory or model need not meet the formal requirements of such; any coherent and plausible conceptual scheme that fulfills this role should suffice. On the other hand, "hypotheses" that are ad hoc or based on hunches should simply be outlawed.

The second circumstance is the converse of the first: *The outcome is not hypothesized but should have been hypothesized (i.e., is hypothesizeable)*. I am sure that most researchers have had the experience of trying to interpret a significant post hoc effect and in the process of doing so realize that there was a sound basis for hypothesizing the effect. (The hypothesis would be the generalized form of a prediction of the effect.) However, I am equally sure that most parapsychologists would not retrospectively change the status of the effect from post hoc to hypothesized (and reap the rewards of doing so) because it looks like cheating. This is a powerful illusion, but an illusion nonetheless. As I noted above, the purpose of a hypothesis test is to provide evidence for or against a theory or model, but to fulfill that role, and for the purported hypothesis to legitimately be designated as such, the relevance of the hypothesis to the theory must have been established. If a proposition meets this test, a full interpretation of the effect requires that it be identified as a hypothesis; otherwise, the support that the confirmation of the hypothesis provides for the theory is obscured. Of course, it is the responsibility of the researcher to justify the reclassification in the Discussion section of the report, and referees can decide whether the author has succeeded. On the other hand, the argument against reclassification is based on the premise that a hypothesizeable effect should only be hypothesized if the researcher was astute enough to recognize that it was hypothesizeable before the study was conducted. This is clearly nonsensical. So my first heretical proposition is that demonstrably hypothesizeable post hoc effects not only can be , but should be, retrospectively reclassified as hypothesized if the researcher becomes aware of its hypothesizeability.

The practical consequences of adhering to my first heretical proposal is markedly reduced by adherence to my second. A major reason why a researcher would want a proposition to be classified as a hypothesis is that the criteria that the prediction(s) derived from it must meet for statistical significance to be claimed are more generous. There are generally two such criteria: (a) a one-tailed rather than a two-tailed significance test, and (b) waiving of the requirement for replication or a multiple-analysis correction. My proposal is that the significance criteria for a hypothesis test should be the same as for a post hoc test, namely, a two-tailed test and a multiple analysis correction or replication.

I have two arguments for my proposal. First, to confirm a hypothesis, a significant effect must be shown to be "real," and this latter determination should be made irrespective of whether the higher-level proposition is classified as a hypothesis. To do otherwise is to assume what is at issue: that is, that classification of the proposition as a hypothesis is proper.

My second argument is similar to my objections to the use of Bayesian statistics, which I presented in a previous editorial (Palmer, 2011). What Bayesian statistics essentially does is to allow a more liberal criterion to be applied in assessing whether an effect is real if it can be shown that the effect has a high a priori probability of being real. In practice, the a priori probability is usually at least partly determined by whether the overarching theory or hypothesis is consistent with the "established" theory of relevance. Of course, this is the very standard that psi doesn't meet, and, as I argued in the editorial, there are solid grounds for maintaining that a priori probabilities, including those based on theory, should have no influence whatsoever on how we determine the reality of an effect, and therefore, we shouldn't use analysis methods that presuppose that the influence should be something other than zero. The case of hypothesis testing is similar, in that an effect is given an easier path to confirmation if the confirmation is consistent with the theory that the hypothesis is derived from.

The one exception to my proscription of one-tailed tests is when the "hypothesis" to be tested is a replication of a previous finding. The reason is that a significant effect in the opposite direction cancels out the original result leading to the conclusion that the effect is not real, the same conclusion one would draw if the replication outcome were nonsignificant. This is not the case for other hypothesis tests. I also maintain that replications are the only hypothesis tests that should be considered "confirmatory" in the sense this term is used by Kennedy (2016).

One-tailed tests are also appropriate for meta-analyses, at least insofar as they be can be construed as a test of the replicability of a previous finding or set of findings, which I think is almost always the case in practice. I agree with Kennedy (2016) that they should be "prospective." A particularly important question in this connection is how close the methodology of the replication needs to be to that of the original study to qualify for inclusion in the meta-analysis, and how uniform in methodology the original studies need to be for them to qualify as targets. I take a more liberal view on this matter than I believe Kennedy does, but I don't have an argument for this preference. However, evidence from the meta-analysis of Bem, Palmer, and Broughton (2001) that only studies that closely followed the methodology of the PRL autoganzfeld series were collectively significant suggests that a more conservative approach might be a better tactic. Finally, I should note my reservations with using the Stouffer *Z* as a measure of replication (Palmer, 2013).

Bierman et al. (2016)

Bierman chose not to reply to my editorial in the last issue of the *JP* (Palmer, 2016) with a Letter to the Editor but he did reply to me privately (D. Bierman, personal communication, August 4, 2016). He makes some dubious statements in that letter that suggest to me I need to expand on a main point of the editorial, especially because he is likely to circulate these statements privately.

The key point in Bierman's letter to me is that Bierman, Spottiswoode, and Bijl (*BSB*; 2016) were not making any claims about whether QRPs in fact existed in the database. Instead the purpose of the analysis was to say what the effect on the significance of the database would be if one were to assume a certain percentage of QRP studies in the database. As he expressed it: "Our conclusion was that assuming that parapsychologists did behave like main stream colleagues we could 'explain' a large fraction of the effect

size reported in that particular meta-analysis. Our conclusion was not a) the parapsychologists are as bad as the main stream experimenters (that was an assumption); b) experimenter X used QRP Y." In other words, it's all hypothetical.

If that's all it is, the whole exercise was a monumental waste of time (and journal space), but the whole point of my editorial was to demonstrate why this is not in fact the case. I would like to add a few additional observations. First, Bierman's assertion that his conclusion is strictly hypothetical is clearly refuted in the abstract of the BSB paper: "We conclude that the very significant probability cited by the Ganzfeld meta-analysis is likely inflated by QRPs, though results are still significant ($p = 0.003$) with QRPs" (Bierman et al., 2016). This does not describe a hypothetical, "what if" situation. Albeit it refers to a likelihood, but the likelihood is of something real.

Second, it's very telling that in their article BSB never explicitly deny insinuating that there actually were a nontrivial numbers of fraudulent QRPs in the database and that QRPs were committed in particular studies. They would have to be imbeciles not to recognize that such inferences by the reader are possible (even likely), and given the seriousness of the potential charges, if their motives really were benign they would have bent over backwards to make this denial clear to the reader.

Bierman's claim in his letter that "Our conclusion was not . . . experimenter X used QRP Y" is also problematic. It is explicit in their subsection on fraud that two particular studies were chosen for the QRP designation, to the point that with the aid of their disappearing supplement file I was able to identify who the author was. All a reader has to do to identify which authors committed a misclassification QRP (e.g., optional stopping or extension) is check the ganzfeld literature for studies with nonround *N*s. Of course, Bierman is correct that none of these attributions were "conclusions." They instead were insinuations, and as I noted in the previous editorial, in my mind insinuations are worse than flat-out allegations. Indeed, this was the basis of my comparing the BSB article to the writings of Hansel.

It is important to recognize that the fraud insinuations do not generally appear in the description of the meta-analysis per se or of its results. For instance, the unjustified inference of misclassification from a nonround sample size informed only the estimate of the proportion of studies in which that QRP occurred. BSB try (unsuccessfully) to justify this inference by pointing out that that the percentage of nonround studies in the ganzfeld database is similar to the estimate obtained by John, Lowenstein, and Prelec (2012) for misclassification QRPs in psychology experiments.

Finally, it should be noted that the BSB paper was published in a psychology journal rather than a parapsychology journal, which means that its target audience was mainstream psychologists. I don't need a crystal ball to tell me that these readers reached the conclusion expressed in the abstract, which I suspect is as far as many of them (and especially the media) went.

References

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

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

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

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.

Kennedy, J. E. (2016). Is the methodological revolution in psychology over or just beginning? *Journal of Parapsychology*, *80*, 156–168.

Palmer, J. (2011). On Bem and Bayes [Editorial]. *Journal of Parapsychology*, *75*, 179–184.

Palmer, J. (2013). JP publication policy: Statistical issues [Editorial]. *Journal of Parapsychology*, *77*, 5–8.

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