EDITORIAL

JP Publication Policy: Statistical Issues

Before I proceed to the topic expressed in the title, I want to officially inform readers of two broader changes in the *JP*. First, our electronic edition, in the past available only to members of the Parapsychological Association (PA) through the PA website, will now be available to all subscribers through the Rhine Research Center's website (www.rhine.org), along with electronic copies of back issues. The second change, already obvious to those of you who are reading our print edition, is an increase in the page size from 6 x 9 to 8.5 x 11 inches. This is the first time the *JP* has ever appeared in this larger form. There are two reasons for the change. First, we save substantially on our printing costs. Second, it is easier for my Managing Editor, Dave Roberts, who also is in charge of producing our finished product, to handle the large tables and figures authors frequently submit.

On to Statistics

The PA recently adopted some radical new guidelines for the presentation of the results of statistical analyses in research papers submitted for its annual convention. Were it not for a technicality, adherence to these guidelines would have been required for papers submitted for the 2013 convention. I do not entirely agree with these guidelines, which seem to be based on trends that are gaining currency in mainstream psychology (Cumming, 2012). Nonetheless, they led me to conclude that the *JP* also should have statistical guidelines. I decided to present these guidelines in this Editorial, in addition to separate guidelines for authors, for two reasons. First, what I have to say should be of interest to all readers who would like to know how we determine statistically whether a psi effect is real. Second, I will be presenting my personal opinions in addition to the publication requirements; they are not always the same. Of course, it would be practically impossible to discuss every statistical issue an author might confront, so I will restrict myself to the issues I consider most important and controversial in contemporary parapsychology.

Effect Sizes

Effect sizes (ESs) are standardized estimates of the magnitude of an effect. Much has been made recently about the desirability of including them in experimental reports, even to the point of suggesting that they are a better measure of a study's success than the traditional *p* value. Essentially the same viewpoint has been espoused by parapsychology's most prominent statistician, Jessica Utts (1988). In my opinion, effect size is the better measure of success only if everyone agrees that an effect is genuine and the real question is whether it is big enough to be of practical value. This circumstance often arises in medicine where, for example, you are concerned with the practical effectiveness of a new drug. If, for example, the drug is shown to be effective in more patients than placebo to a statistically significant degree because of a large sample size, there's nothing to get excited about if the success rate of the drug is 5% and the success rate of the placebo is 3%. Here, size clearly matters. (However, in this and similar examples, the raw percentage would seem to be more informative than the ES; the ES seems preferable only if there is need for a standardized measure, as is often the case in meta-analysis). However, in parapsychology, the question is almost always whether the effect is real, not how big it is if it is real. To answer the "reality" question, it is the *p* value that should be consulted. There are several ways to legitimately describe a *p* value. The most relevant description to parapsychology I have found appears in a professionally written essay in Wikipedia (2013): "[a *p* value] is a measure of how likely the data is to have occurred by chance, assuming the null hypothesis is true." Control issues notwithstanding, the (non)likelihood of chance is the basis on which we declare that a psi effect is real, for valid reasons that should be obvious.

Although I question the usefulness of ESs in most parapsychology papers, authors are free to use them and the corresponding confidence intervals (CIs), provided the *p* value is reported as well and the effect is statistically significant. Why the latter exception? When you say an effect is not statistically significant, you are in effect claiming that the ES is zero in the population. To follow this with a nonzero ES contradicts the implication of the *p* value. Although one might reply that the citation is justified to honor the possibility of a Type I error, the combination is close enough to self-contradictory to be objectionable. However, as stipulated in the current Instructions for Authors, authors must continue to report other descriptive statistics, such as means and standard deviations, regardless of statistical significance. Finally, and most importantly, I will never allow an ES or other magnitude estimate to serve as the sole justification for a claim that a psi effect is real.

The Multiple Analysis Problem

By far the most common statistical criticism I get from my referees concerns "multiple analyses." The point is that when you perform a large number of analyses, you expect some of these to be significant by chance. Therefore, the criterion for significance for any one of these multiple effects must be made more conservative to correct the problem.

To begin with, my publication policy from now on will be that any effect that is statistically significant by the traditional criterion must be claimed as tentative unless the researcher does something acceptable to justify a stronger claim. I refer to such statements as "disclaimers." They can apply to individual effects, but in most cases a single blanket statement can be made covering most or all of the effects reported for the experiment.

Adjusting the alpha level. So what can be done to avoid the disclaimer? The most common solution to the multiple analysis problem is to make the criterion for statistical significance more rigorous by applying a multiple analysis correction. The most common of these corrections in parapsychology is the Bonferroni adjustment. You simply take the original significance criterion, usually .05 in psychology and parapsychology, and divide it by the number of analyses you need to adjust for (which I call the "base *N*"), and this becomes the new criterion (alpha level) that the effect must meet. For example, if there are 10 base *N* analyses, you divide .05 by 10 and get .005. Any effect requiring the adjustment must be $p < .005$ to be considered significant.

I have serious concerns about these multiple analysis corrections. First, there seems to be no consensus as to whether they should be applied only to unplanned post hoc effects (data snooping), or also planned post hoc effects, or even hypothesized effects. Similarly, there is no sound justication for any particular base-*N* criterion; it's rather arbitrary. It used to be that all you were supposed to correct for were paired comparisons inside an ANOVA table using a Scheffé test or such. Then it came along that you had to correct for many more of the analyses in the study, but there seems to be disagreement on how many more. For example, can you exclude hypothesized effects or even planned post hoc effects? Most post hoc analyses in my own research are unplanned attempts to understand some other significant effect in the study. I would argue that these secondary analyses shouldn't count in the base *N* because they are yoked to a primary effect. However, I'm sure this could be argued either way, and I doubt a multiple analysis correction aficionado would have much sympathy for my plea. So let's agree to elevate the base *N* to include all the analyses in the study. But why stop there? It seems to me that if you follow this multiple analysis reasoning to its logical conclusion, the proper base *N* (at a minimum) should be all the analyses ever conducted in the social/behavioral/neurophysiological sciences, which of course would guarantee that no analysis could ever be claimed as significant.

However, my main objection to multiple analysis corrections is that the whole notion that the de facto objective likelihood of an effect being real is influenced by how many other analyses an investigator decides to conduct is absurd on its face (unless, of course, one assumes a paranormal cause!) Demanding the correction could also discourage investigators from conducting exploratory analyses as a way to protect the significance of the more important analyses. This is especially a problem in parapsychology because the great majority of our studies are primarily exploratory, and we need good hypothesis generation. They also markedly increase the Type II error rate. Finally, I have a particular problem with the Bonferroni because it assumes that the analyses are independent; this assumption usually is grossly violated in practice.

So, my policy on multiple analysis corrections is the following. Authors are free (but not required) to apply them to however many analyses they wish, so long as they specify the base *N*(s) and describe any classes of analyses that are excluded. However, these corrections cannot be used to avoid the disclaimer.

Replication. So, what's the alternative? My position is that to avoid the disclaimer, any initially discovered significant effect must be successfully replicated, ideally (but not necessarily) by an independent investigator. An important virtue that replication supplies, which a huge *p* value for the original effect cannot supply, is evidence of generality or robustness, that the effect is not beholden to a particular sample or set of experimental circumstances. This is particularly true in parapsychology, where the effects are subtle or elusive, and we don't have a very good idea of all the factors that can influence an experimental outcome. Even a replication by the same investigator is different than the original study, if for no other reason than the experimenter will approach it with a different attitude, caused by the success of the original study. Moreover, the sample of participants can be unexpectedly different in some crucial respect (see, e.g., Stanford & Frank, 1991). Because of this principle of generality, I would be more impressed by the combined statistical significance of 10 distinct parapsychology studies with 50 participants each than a significant result from a single study with 500 participants; multiple studies are better.

So how do we define a successful replication? My preference is a traditional criterion, statistical significance at the .05 level, one-tailed. However, I sense that there is a lack of agreement in our field about this criterion, so I will *allow* a more liberal criterion in the *JP*. At this time, the most liberal alternative I am willing to accept is combined significance of the original and replicating study using the Stouffer *Z*, as illustrated in the study by Dalkvist in this issue. What I will not allow is successful replication to be claimed if the effect size (ES) of the replication study falls within the 95% CI of the effect size of the original study. Although application of this criterion might be defensible in other fields, where one is seeking a reliable point estimate for the magnitude of an effect, in our field it has the disastrous consequence of allowing an investigator to claim evidence for psi regardless of the outcome of the analysis: If the Study 2 ES falls within the Study 1 CI, the investigator claims a successful replication; if it falls outside this CI, the investigator claims a statistically significant difference between the two studies. Imagine the fallout if our critics ever caught on to this!

In my opinion, replication should be required for any effect regardless of (a) whether it was unplanned post hoc, planned post hoc, or hypothesized, (b) the *p* value (adjusted or unadjusted), and (c) the sample size. I think I've explained my reasons for (b) and (c) adequately above, but (a) needs more comment, particularly with regard to hypothesized effects. The claim that hypothesized effects should be treated more leniently than other effects is the flip side of the statement routinely uttered by critics of parapsychology that "extraordinary claims require extraordinary proof." Many years ago, I wrote a paper attacking this claim, arguing that the criteria an effect must meet to be accepted as a scientific fact should be uniform, at least within a broad scientific domain such as psychology (including parapsychology; Palmer, 1987). To give preference to findings because they are consistent with some hypothesis or theory tends over time to unfairly bias the literature in favor of that hypothesis or theory. In the present case, this problem would manifest by declaring as real an unreplicated effect hypothesized by, and thus consistent with, some theory, while labeling the same finding, if it did not support (and indeed might be inconsistent with) the theory, as only tentative.

However, for *JP* publication policy, I will again at least partly bow to what I perceive as the more commonly held view within parapsychology and adopt a criterion more liberal than my personal preference. Thus, I will allow significant hypothesized effects, but not significant post hoc effects (planned or unplanned), to escape the disclaimer, which should read something like "The finding(s) is (are) tentative pending replication."

Registration of Experiments

In a recent letter to the *JP*, Caroline Watt (2012) announced that the Koestler Parapsychology Unit of the University of Edinburgh was setting up a registry to which researchers are invited to submit their hypothesized and planned analyses in advance of data collection. The purpose of the registration is to provide concrete evidence that these analyses are "planned prior to conducting the experiment" (p. 403). Insofar as authors abide by my preference that nonconfirmatory hypothesized and planned analyses be treated the same way statistically as unplanned analyses, they need not submit such analyses to the registry. On the other hand, I think the registry is a good idea for other analyses. My only caveat is that I think the registry would have more credibility among researchers and achieve maximum usage if it were sponsored by the PA rather than any individual laboratory. I would only consider registration by the previously defined "unencumbered" researchers to be a requirement for publication in the *JP* if the same policy were agreed to by the editors of all the major journals in the field.

Power Analyses

Statistical power is defined as the (a priori) probability that a false null hypothesis will be rejected. Its primary value is to alert investigators how large their sample must be if they hope to have a decent chance to successfully replicate a previously obtained effect. This is a sobering exercise, because researchers are likely to discover that the needed *N* is larger than they intuitively expect and can readily obtain in practice. My policy for the *JP* will be to recommend, but not require, that the power statistic be reported in cases of replication. However, I consider power analysis potentially problematic for effects other than replications. The reason is the need to insert an ES estimate in the power analysis formula. In the case of a replication, the solution is straightforward; you insert the ES from the original study. However, for other effects, particularly exploratory ones, there is often no sound basis for estimating an ES. I ran into this problem myself a few years ago when I submitted a psychology article to a mainstream journal. The editor originally insisted that I publish a power statistic, even though the effect was unprecedented in the literature and I didn't have a clue how to estimate an ES for the analysis. Fortunately for me, the editor relented and the paper was published without a power statistic (Palmer, Mohr, Krummenacher, & Brugger, 2007).

For the *JP*, if authors want to report a power statistic in nonreplication cases, they must provide a sound justification for the ES they insert in the power analysis formula. Moreover, authors must use extreme caution in attributing the failure of an effect to reach significance to low power, even if the ES is large. With low *N*s, effect sizes can be highly unstable, and a high one could easily be the luck of the draw. The burden of proof falls on the researcher to conduct a large enough experiment to obtain a significant *p* value.

Conclusion

One of my perks for being a journal editor is that I get to write editorials in which I can say pretty much anything I want without having to worry about my remarks being vetoed by some referee. The current editorial is of course a case in point. However, readers are free to express their disagreement with my views in a Letter to the Editor, which I will publish so long as it's in good taste. Even if there are no Letters, I hope my remarks will provoke thought and discussion about these matters among the readership. That's always a healthy development.

References

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Palmer, J. (1987). Dulling Occam's Razor: The role of coherence in assessing scientific knowledge claims. *European Journal of Parapsychology, 7*, 73–82.

Palmer, J., Mohr, C., Krummenacher, P., & Brugger, P. (2007). Implicit learning of sequential bias in a guessing task: Failure to demonstrate effect of dopamine administration and paranormal belief. *Consciousness & Cognition, 16*, 498–506.

Stanford, R. G., & Frank, S. (1991). Prediction of ganzfeld-ESP task performance from session-based verbal indicators of psychological function: A second study. *Journal of Parapsychology, 55*, 349–376.

Utts, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology, 52*, 305–320.

Watt, C. (2012). [Letter]. *Journal of Parapsychology, 76*, 403.

Wikipedia. (2013). *p-value.* Retrieved from http://en.wikipedia.org/wiki/P-value

John Palmer