Journal of Parapsychology
2019, Vol. 83, No. 2, 117-119

http://doi.org/10.30891/jopar.2019.02.01

Editorial: Errors of the Third Kind [1]

Etzel Cardeña

Just like tidal unveilings of flotsam, science discovers “anew” errors of the third kind, Schlaiffer’s term for the misuse and misinterpretation of statistical procedures besides the Type I and II errors (Schlaiffer, 1959, in Bakan, 1966). In this issue, Jessica Utts and Patrizio Tressoldi mention the alarm cries of a “credibility revolution” or the “replicability crisis,” to which could be added the bugaboo of “questionable research practices” (QRP). Have scientists become more insightful recently about the problems of mindless applications of statistical and research procedures? No, they are just reiterating ideas that have been around for more than a half century. As I mentioned in a previous Editorial (Cardeña, 2017), paraphrasing the famous quotation by Jorge Santayana (1905), science is condemned to repeat what it cannot remember.

Let me take a few “recent” ideas and verify whether they had already been discussed in a 1966 paper by my former mentor David Bakan: File drawer effect because most journals will not publish failures to replicate? Check! Misunderstanding of the real meaning of the *p* statistic, with some authors inferring a lot more from it than is warranted? Check! Selecting one of multiple analyses without reporting the others? Check! And there are more checks, but I will not tire the readers and instead recommend that they read Bakan’s insightful work. He also made a clear distinction between general and aggregate functions, the first one referring to values that are true for all members of the group, whereas the second (e.g., measures of central tendency) refers to an aggregate of values and may reflect few if any of the actual values of the members of that group (Bakan, 1967). This is an essential point that clarifies why the sciences of living, sentient, (and, in the case of humans, historical) beings, typically based on aggregate statistics, will never approach the precision of the exact sciences, which includes fully generalizable results in many areas. This issue also partly explains why the discourse of a replication “crisis” in psychology has been exaggerated and is not quite coherent (for a paper treating this problem and recommending multiple measures of the same individual in different contexts, an approach that psi research should do well to adopt whenever possible, see Epstein, 1980; for a more recent discussion of the exaggeration of the problem, see Barrett, 2015). Bakan also anticipated what I think is becoming an increasing problem, namely the use of large online surveys that produce significant *p* values (the *p* statistic is very sensitive to the size of *N*) no matter how theoretically and practically negligible (and probably unreliable) those differences might be.

Which brings me to the recent and authoritative criticisms of the typical (mis)use of significance values. The American Statistical Association (ASA) has developed six principles to clarify *p*-values, which given their importance I transcribe:

[1] Send correspondence to: Etzel Cardeña, Ph. D., CERCAP. Department of Psychology, Allhelgona kyrkogata 16a, Lund 22100, Sweden. etzel.cardena@psy.lu.se

1. "*P*- values can indicate how incompatible the data are with a specified statistical model...
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone...
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold...
4. Proper inference requires full reporting and transparency...
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result...
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis." (Wasserstein, 2016, pp. 131-132)

The ASA concludes that "Good statistical practice... emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean" (Wasserstein, 2016, p. 132). Instead of mindlessly adopting a somewhat arbitrary value for what is/is not of scientific significance, the ASA recommends fully grappling with the data, as well as accepting a level of uncertainty and ambiguity in the scientific process. Or, as wittily put by the eminent statistical psychologist Robert Rosenthal in various presentations and publications, "surely God[ess] loves the .06 as much as the .05" (e.g., Rosnow & Rosenthal, 1989, p. 1277).

Just some weeks ago I witnessed how widespread is the problem with the over-reliance and misinterpretation of *p*. A young psychologist was giving a presentation on how a certain group of parents was "significantly" more likely to produce psychological problems in their offspring than another. S/he had a slide with a graph showing the distribution of the scores in question. Given that the distributions of both groups overlapped considerably, I asked him/her about the effect size of the difference. She had no answer other than that the difference was "significant," so I asked then how clinically/practically relevant was the difference between the two groups, and s/he again had no response. Of course, other researchers have been using meta-analytical, Bayesian, and other approaches as alternatives to the mindless use of the null hypothesis significance testing approach.

I was fortunate to learn from Bakan to reflect critically about *p* values and other scientific automaticities, and because of that I started using effect sizes before they became fashionable. Thanks to him I also became aware of the Bayesian approach decades before it was better known. This general awareness, aided by the expertise of Utts and Tressoldi (2015) informed the statistical guidelines for this journal, which I will ask authors and reviewers to enforce more strongly. I am not ready to proscribe the use of "significance" language, as some (Hurlburt, Levine, & Utts, 2019) have done, but will require from authors a justification of why any result with a $p < .05$ is trustworthy and important, as well as asking them to discuss relevant results even with a larger *p*. Also, following the current trend in academic journals, from 2020 onwards the *Journal of Parapsychology* will follow a hybrid open access model in which authors will be able to keep the copyright of their contributions provided they pay a fee (for details see the "Guidelines for Authors" in this issue).

I hope that David Bakan can somehow know that indirectly he continues to help keep the epistemological beach clean from "errors of the third kind."

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437. doi.org/10.1037/h0020412

Bakan, D. (1967) *On method: Toward a reconstruction of psychological investigation.* San Francisco, CA: Jossey-Bass.

Barrett, L. F. (2015, September 1). *Psychology is not in crisis.* Retrieved from www.nytimes.com/2015/09/01/opinion/psychology-is-notin-crisis.html?_r_0

Cardeña, E. (2017). *On scientific amnesia. Journal of Parapsychology, 81,* 104-105.

Epstein, S. (1890). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790-806. doi.org/10.1037/0003-066X.35.9.790

Hurlburt, S. H., Levine, R. A., & Utts, J. (2019). Coup de grâce for a tough old bull: "Statistically significant" expires. *The American Statistician, 73,* 352-357. doi.org/10.1080/00031305.2019.1543616

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284. dx.doi.org/10.1037/0003-066X.44.10.1276

Santayana, G. (1905). Reason in common sense. New York: Charles Scribner's Son. Retrievable from www.gutenberg.org/files/15000/15000-h/vol1.html#CHAPTER_I_THE_BIRTH_OF_REASON

Schlaiffer, R. (1959). *Probability and statistics for business decisions.* New York: McGraw-Hill.

Utts, J., & Tressoldi, P (2015). Statistical guidelines for empirical studies. In E. Cardeña, J. Palmer, J., & D. Marcusson-Clavertz, D. (2015). *Parapsychology: A handbook for the 21*st century (pp. 83-93). Jefferson, NC: McFarland.

Wasserstein, R. L. (2016). ASA statement on statistical significance and *p*-values. *The American Statistician, 70,* 131-133. doi.org/10.1080/00031305.2016.1154108

p.s.: See also in the Correspondence section a letter by Caroline Watt and Jim Kennedy on current discussions of confirmatory versus exploratory analyses.

p.p.s.: This issue contains the abstracts of the last meeting of the Parapsychological Association. Because many of them were quite a bit longer, I shortened them and did some light copy-editing for grammatical and other problems. The email of the address of the first authors is included for those wanting more details of their work.