“If your experiment needs statistics, you ought to have done a better experiment.”
Ernest Rutherford

Still too many significant findings are false (and nonsignificant findings true):
A comment on Benjamin et al. (2017), Redefine statistical significance
R. Hans Phaf

Does the redefinition of significance level by Benjamin and colleagues (2017) turn a majority of false positives (i.e., significant but “false” findings; Ioannidis, 2005) into a minority? For a typical power of 0.40 (Cohen, 1962; Sedlmeier & Gigerenzer, 1989), and with 10 times more false than true hypotheses (prior odds in experimental psychology assumed by Benjamin et al.) it does. According to the formulae of Ioannidis, the redefinition raises the positive predictive value (probability of significant findings being “true”, PPV) from an uncomfortable 0.44 to a reasonable 0.89. Note that these calculations assume no bias or questionable research practices (QRP), and that false-positive rates worsen further with QRPs (and multiple independent testing). Of course, this hinges on what counts as a false hypothesis. Given sufficient power, almost all hypotheses, even those predicting an infinitesimally small difference, may be true (Cohen, 1990). If all hypotheses are true, the false-positive rate becomes zero. Most likely in psychology, with its more coarsely grained criteria for an effect, false hypotheses outnumber true hypotheses by much more than a factor ten. The factor may not be as large as in genomic research (100,000 according to Ioannidis, 2007), but even Johnson, Payne, Wang, Asher, and Mandal (2017) concede that it is probably larger than 10 in experimental psychology. In a simplistic scenario of 1000 independent and 1000 dependent variables with two types of relationships (i.e., increasing, decreasing), and with each independent variable having a bearing on 1% of the dependent variables, this factor rises to 200. When the factor exceeds 80, false positives gain the majority again, even with a significance level of 0.005. From a false-positive perspective the significance level will need to be reduced much further to achieve acceptable confidence in significant results. If psychology anywhere approaches genomics research, it would need a significance level of at least 5 x 10^{-6}.

Obviously with such false-positive rates, it will be hard to replicate a majority of published findings in psychology (Open Science Collaboration, 2015). Still due to the also very high false-negative rates (nonsignificant but “true” findings), it cannot be excluded that ALL of the OSF nonreplications represent false negatives (Hartgerink, Wicherts, & van Assen, 2017). In addition, of 6,951 articles from top psychology journals reporting a nonsignificant result, Hartgerink et al. calculated 66.7% to contain at least one false negative. The tendency to not publish nonsignificant results can only raise the estimate of the incidence of false negatives. Cohen (1962; see also Sedlmeier & Gigerenzer, 1989) surveyed papers published in the Journal of Abnormal and Social Psychology at that time and concluded that small-effect studies had a false-negative rate of 0.82, medium-effect studies of 0.52, and large-effect studies of 0.17. As has been carefully argued by Fiedler, Kutzner, and Krueger (2012), high false-negative rates may be much more detrimental to science than high false-positive rates. The former may prematurely close the door on “true”, innovative hypotheses, whereas the latter may be subject to continued research and eventually be corrected, if “false”. Innovative hypotheses, by their nature, are plagued by a lack of knowledge of the best task settings, stimuli, instructions, and context conditions (cf. Fiedler et al., 2012), so that they will generally tend to yield smaller effects, and more false negatives, than well-established hypotheses. From a false-negative perspective, it makes more sense to increase, rather than reduce, the significance level for novel findings, if at all a meaningful definition of such novel findings can be formulated.

To many of us in experimental psychology, who have been trained to rely heavily on significance testing, the high false-positive rates (and also the high false-negative rates) may have come as a shock. Benjamin et al. (2017) even call it “unsettling”, but are apparently prepared to put up with
false-positive rates below 10%, assuming prior odds of 1:10. Lower prior odds may make these rates intolerable even to them. It is unclear what level of false negatives they would be able to tolerate, or whether they care at all about false negatives. The overreliance on significance testing has created the illusion that a single, significant, and seemingly novel, finding provides conclusive evidence and can stand on its own, so that it does not need to be placed in the context of other findings or theories. Too often significance then serves as a stop criterion for theoretical analysis (Phaf, 2016). This has led to (a) the wheel being reinvented over and over again, (b) opposite, but both significant, research claims being published shortly after one another (i.e., the Proteus phenomenon; Ioannidis & Trikalinos, 2005), and worse still (c) a non-incremental development of theory with hypotheses remaining highly popular after they have been rather firmly refuted (cf, Begley & Ioannidis, 2015). Significance is not only practically inconclusive due to the high false-positive and false-negative rates, but it is also theoretically vacuous. Significance would at best only indicate what the effect is not, and leaves open an uncountable infinity of alternative hypotheses. Instead of collecting conclusive facts, significance testing has thus left us with a great deal of fallacious “facts” and rejected “truths” and is clearly responsible for the highly confused state psychology appears to be in at the moment. Rather than trying to salvage the fatally flawed practice of null-hypothesis significance testing (Cumming, 2014), psychology should improve its attempts at theoretical analysis and integration.

References