Publsihed in *Educational and Psychological Measurement*, 2017, *77*(4), 631-662

Thou shalt not bear false witness against null hypothesis significance testing

Miguel A. García-Pérez
Departamento de Metodología, Facultad de Psicología, Universidad Complutense,
Campus de Somosaguas, 28223 Madrid, Spain. E-mail: miguel@psi.ucm.es

Abstract

Null hypothesis significance testing (NHST) has been the subject of debate for decades and alternative approaches to data analysis have been proposed. This paper addresses this debate from the perspective of scientific inquiry and inference. Inference is an inverse problem and application of statistical methods cannot reveal whether effects exist or whether they are empirically meaningful. Hence, raising conclusions from the outcomes of statistical analyses is subject to limitations. NHST has been criticized for its misuse and the misconstruction of its outcomes, also stressing its inability to meet expectations that it was never designed to fulfil. Ironically, alternatives to NHST are identical in these respects, something that has been overlooked in their presentation. Three of those alternatives are discussed here (estimation via confidence intervals and effect sizes, quantification of evidence via Bayes factors, and mere reporting of descriptive statistics). None of them offers a solution to the problems that NHST is purported to have, all of them are susceptible to misuse and misinterpretation, and some bring around their own problems (e.g., Bayes factors have a one-to-one correspondence with *p*-values, but they are entirely deprived of an inferential framework). Those alternatives also fail to cover a broad area of inference not involving distributional parameters, where NHST procedures remain the only (and suitable) option. Like knives or axes, NHST is not inherently evil; only misuse and misinterpretation of its outcomes needs to be eradicated.

CIS Keywords: Estimation; Significance testing; Inverse problem; Bayes factor; Goodness of fit

Thou shalt not bear false witness against null hypothesis significance testing

"The difficult I'll do right now
The impossible will take a little while"
– From *Crazy He Calls Me*; Sidney Keith "Bob" Russell, 1949 –

The goal of psychology as a science is to gain knowledge and understanding of all aspects of human behavior, from the operation of perceptual and cognitive processes, through the structure of personality and systems of attitudes, interests, or beliefs, to interpersonal behavior at large. Accomplishing this goal requires meticulous design of studies and a proper analysis of the data that can answer the posed research questions. Researchers expect the data to be unambiguous so as to be able to draw conclusions that convey the intended step forward. Sometimes the resultant pieces of knowledge take the form of meaningful evidence on the prevalence of certain patterns of overt behaviors or opinions (in observational or survey studies), on the strength of association between variables (in correlational studies), on differential characteristics among groups (in *ex post facto* studies), or on causal relations between variables (in experimental studies). In other cases, knowledge of a functional nature comes from empirical evidence consistent with the predictions of quantitative models that describe observed data as the result of the operation of hypothetical processes characterized by latent parameters.

With the exception of case studies, research results are never reported with the simple purpose of describing the current data while acknowledging that other studies will surely yield different outcomes. Instead, researchers want to think of their results as solid pieces of evidence in some direction. This means that research has an essentially inferential purpose: Studies are designed in particular ways because some aspects of the data thus collected can be regarded as dependable indicators of an underlying reality rather than simply viewed as an incidental description of the current sample. A principled inferential process then extracts an interpretation that seems plausible, although observed data are usually compatible with a relatively large set of alternative interpretations. Even if the study was properly designed and dependable research methods were used, observed data are inevitably tainted by sampling error, measurement error, and individual differences. Research results are most informative when conclusions of some generality can be drawn from the observed data, which requires an analysis capable of detecting systematic patterns unlikely to be the result of variability caused by those sources of error. This inferential process calls for objective methods of data analysis, where "objective" is only antonymous to "subjective" but not synonymous with "revealing the truth." In an admittedly oversimplified account at this point, the use of such methods may result in confidence intervals for distributional parameters (e.g., proportions, means, correlations, ..., or differences between them) or in a decision regarding the tenability of some experimental hypothesis that was translated into a statistical hypothesis involving distributional parameters. Despite the objective inferential process, its outcome is heavily dependent on the tainted observed data and, hence, the final conclusion may be erroneous in comparison with the always unreachable truth. In these conditions, one can only ask that the objective inferential process is well calibrated so that the frequency with which it leads to erroneous conclusions is known and under the researcher's control. One would also ask that researchers are aware both of these limitations and of their implications, so that the scope of the research conclusions presented in a paper is neither misreported by the authors nor misconstrued by the readers.

How statistical analyses should be conducted and their outcomes be interpreted has been the subject of heated debate for some decades, mostly focusing on the widespread practice of routinely subjecting the data to null hypothesis significance testing (NHST). The main point in question was whether this practice fosters or impedes progress in psychology. An exhaustive and unparalleled analysis of the controversy was conducted by Nickerson (2000), who compiled a large number of the criticisms that had contributed over the

years to questioning the adequacy of NHST. The criticisms come down to false beliefs, misconstructions, misinterpretations, and unfulfilled expectations. Put another way, they only reflect poor understanding and widespread misuse (or plain abuse) of NHST on the part of researchers. In a humanly but inexplicable move, misuse and misinterpretation of NHST soon turned to be regarded as a problem of NHST itself, a methodology that is now purported to suffer from "problems" or "fatal flaws" and criticized for not allowing the type of inferences that researchers seek (e.g., Branch, 2014; Cumming, 2014; Gill, 1999; Goodman, 1999a; Krueger, 2001; Lambdin, 2012; Schneider, 2015; Wagenmakers, 2007). Accordingly, alternative methods of analysis that presumably overcome these "problems" have been developed and advocated.

This paper will leave misuse and misinterpretation of NHST aside. These issues have been extensively discussed over the years and corrective action is well defined (though rarely implemented), so there is no reason to discuss it further. Instead, the main focus of this paper is a discussion of the presumed benefits that alternative methodologies bring in comparison to what proper NHST practices would have brought. The approach taken in this paper thus contrasts with analyses in which the outcomes of such alternatives have been confronted with those of an improper use of NHST.

The paper is organized as follows. First, the inherent limitations of inferential procedures are laid out, which certainly affect NHST but also all of its alternatives. Then, three particular alternatives are discussed in this inferential context, namely, a switch towards estimation via confidence intervals and effect sizes, a switch towards quantification of statistical evidence via Bayes factors, and a switch towards mere reporting of descriptive statistics. The discussion shows that none of these alternatives solve the presumed problems of NHST, also showing that they have their own problems besides being equally susceptible to misuse and misinterpretation. Next, NHST is shown to provide a proper inferential framework for the purpose of assessing the goodness of fit of process models, an area where alternative inferential procedures offer essentially the same solution. Before the final discussion, two problems that are often addressed inadequately via NHST procedures are briefly commented on.

The unfulfillable expectation

Generally speaking, researchers seek evidence in favor of their hypotheses, be they worded positively or negatively. This manifests in the way in which the conclusions of a study are expressed, sometimes making it even into the title of the paper. The conclusion is often stated categorically (e.g., individuals of type X are more Z than individuals of type Y, or performance is higher under condition X than it is under condition Y), as if the study had shown such statement to be incontestable and universally true. In a milder account, Cohen (1994) stated that researchers want to be able to determine the probability that their research hypothesis is true, given the observed data. The interpretation of such proposition is elusive. In principle, the concept of probability implied in it must be understood somewhat loosely because, strictly speaking, the probability that a continuous parameter (e.g., the difference between two means) attains any precise value is well known to be zero, were this regarded as a random variable instead of the unknown fixed value that it really is. But what would it mean that the probability of, say, a female advantage in ability to negotiate is .8? Perhaps that average gender differences of some specified size will be observed in 80% of the studies? That 80% of the females are more successful than males at negotiating? Or maybe that one can be 80% confident that a female advantage really exists? Appealing as Cohen's proposition seems, it lacks a referent for "probability that a hypothesis is true," which becomes a formidable obstacle to estimating what value such probability might have. In any case, NHST certainly cannot lead to categorical statements about the truth of hypotheses, nor can it provide probabilistic statements about the likelihood that a hypothesis is true (whatever this might mean). This is surely discomforting when one seeks certainty (or quantification of uncertainty) and it may be one of the reasons why researchers translate NHST results into claims that they find satisfactory. Thus, the interpretation of statistically significant results is almost always worded as an unreserved confirmation of the

alternative hypothesis (i.e., the means differ, a correlation exists, etc).

NHST was never designed to find out whether or not a hypothesis is true. NHST procedures derive the probability distribution that a test statistic computed from the data would have if the null hypothesis were true, which the researcher then uses as a reference to make a decision to reject or not to reject the null. The null is rejected if the sample value of the test statistic is too extreme under the reference distribution (i.e., if it exceeds some critical value in the appropriate direction). But rejecting the null is not synonymous with proving it wrong (or unlikely), just as not rejecting it is not synonymous with proving it correct (or likely). Such decisions have been purported to be the result of a faulty syllogistic reasoning with probabilistic premises (e.g., Branch, 2014; Cohen, 1994; Gill, 1999; Wagenmakers et al., in press); yet, they only reflect the outcome of a decision rule devised on consideration that any decision made under uncertainty on any grounds will sometimes be inadequate. Null hypotheses may thus be undeservedly rejected or not rejected, but whether the decision was adequate on any particular occasion will always remain unknown: The truth about the hypothesis is always beyond reach. In these circumstances, one would only expect that the decision rule lends itself to an examination that permits quantification of the frequency with which its outcomes will be inadequate, that is, how often application of the decision rule will reject a true hypothesis and how often it will not reject a false hypothesis. This is one of the undeniable and unique strengths of NHST. Asymptotic Type-I and Type-II error rates of NHST procedures can be determined with precision and these hold up fairly accurately in small-sample conditions in the vast majority of cases. In the end, this implies that research can be conducted in a way that the NHST procedures called upon rarely reject (or not reject) a true (or false) hypothesis, which should not be mistaken to imply that rejected nulls are false or that not rejected nulls are true. As discussed earlier, the truth of a null can never be determined.

Admittedly, the critical value used for rejecting or not rejecting a hypothesis is the researcher's choice and, thus, NHST outcomes are amenable to subjective decisions. Initially, critical values for a small set of significance levels were tabulated for routine use although recourse to the $\alpha = .05$ significance level prevailed in practice. This only means that researchers accepted the risk of rejecting the null on 5% of the occasions in which NHST procedures were used to test a null hypothesis that happened to be true, also accepting alongside the risk of not rejecting a false null hypothesis on a percentage of occasions that can be computed considering the size of the sample and a suitable alternative hypothesis (i.e., effect size). When numerical computation became widely available, the *p*-value associated with the sample test statistic could be directly computed and reported instead of making a binary decision by comparison with a tabulated critical value. One way or the other, tagging NHST results as significant only reflects the researcher's principled and transparent decision regarding the criterion for rejection of the null, be it a critical value on the test statistic or a boundary *p*-value. In principle, one could set $\alpha = .3$ and declare significant any result whose *p*-value does not exceed .3. This only means that the researcher intentionally adopts a liberal criterion by which rejecting true null hypotheses is permitted to occur with probability .3. This may be viewed as questionable flexibility (Simmons, Nelson, & Simonsohn, 2011) but this does not prevent other researchers (including journal editors and reviewers) from judging by themselves the meaningfulness or utility of those reportedly significant results, or to assess statistical significance by their own criterion.

In sum, NHST does not allow any conclusion whatsoever about the truth of hypotheses. NHST is only an objective and principled inferential procedure via which hypotheses are rejected or not rejected by application of a well-calibrated decision rule. The probability of inadequate decisions is known, whether the inadequacy is in the form of rejecting a true null hypothesis or in the form of not rejecting a false null hypothesis. On any particular occasion, whether or not the decision was adequate will remain unknown. Little else can be done in the absence of unequivocal empirical signs that a hypothesis is true or false. The inability of NHST to tell whether or not a hypothesis is true may be deemed unsatisfactory, and one may hope for an alternative method of analysis that allows the type of statements that researchers seem more

comfortable with (as judged by the way in which NHST results are typically misreported and misconstrued). Three such alternatives will be discussed in the next section.

It should nevertheless be noted that mounting evidence coming from independent tests of the same hypothesis in replication studies speaks about the empirical tenability of hypotheses, although still not proving them true or false. Consider the issue of gender differences in the ratio of the lengths of the second and fourth digits (2D:4D ratio). Manning, Scutt, Wilson, and Lewis-Jones (1998) reported a "highly significant" difference ($p = .0002$) between males and females in the 2D:4D ratio. This result was worded as "high 2D:4D ratios are characteristic of females." Histograms plotted in their figure 1 for samples of 400 males and 400 females show instead virtually superimposed distributions that only differ in a barely noticeable heavier right tail in the distribution for females and an also barely noticeable heavier left tail in the distribution for males. The effect size was not reported in the paper but it can be computed from the sample sizes and the t statistic, yielding $d_s \approx 0.27$. The study spurred numerous other studies seeking behavioral correlates of these gender differences. Although none of those studies found any remarkable correlation with behavioral measures, all of them concurred in finding analogous gender differences in 2D:4D ratio that were generally statistically significant at the .05 level but with effect sizes that invariably meander around 0.25 only (Hönekopp & Watson, 2010; see also the special section published in 2011 in Vol 51, No 4 of the journal *Personality and Individual Differences*). Gender differences in *average* 2D:4D ratio thus seem established although the effect size is too small to be of any practical significance: The probability that a randomly drawn female will have a higher 2D:4D ratio than a randomly drawn male is only about .6.

Undeniably, this conclusion could also have been reached by an analysis of effect sizes across replication studies, but proper use of NHST does not hamper the conclusion. This example nevertheless illustrates a characteristic of NHST that is also relevant in the assessment of goodness of fit (discussed later in this paper). Specifically, if the null hypothesis (here, equal average 2D:4D ratio in both genders) were true, the decision rule in NHST ensures that significant differences at the α level (for any and all values of α) will be found in 100α% of the replications. Rejecting the null massively across replications increases our confidence that the null is false; rejecting it the stated proportion of times across replications increases our confidence that the null is true. But neither effect sizes nor NHST results can be used to quantify this confidence or to indicate whether believing the null to be false is practically relevant or meaningful. Assessing statistical significance serves a purpose but it is certainly insufficient for an assessment of empirical meaningfulness. This does not make NHST wrong; it only says that researchers should not just stop upon finding statistically significant results without looking further at the data to assess empirical meaningfulness.

Alternatives to NHST

The preceding section discussed that NHST can only reject or not reject a null hypothesis, with no means to discern whether the hypothesis is true or false, or to judge whether the observed departure from the null is empirically meaningful were it authentic. This section discusses three of the various alternatives to NHST that have been proposed, often presented under the perspective that truths may be revealed or, at least, that one can come closer to them than NHST procedures allow (except perhaps for the case of Bayesian analyses, some of whose advocates seem to disdain connections with the empirical truth and emphasize that data be used only to quantify and update subjective beliefs about theoretical positions on the null; see Rouder, 2014).

Estimation (confidence intervals and effect sizes)

Cumming (2014) argued against routine use of NHST and mechanical reporting of statistical significance, proposing a switch towards estimation by the provision of confidence intervals (CIs) and the reporting of effect sizes (ESs). This switch is also aimed at steering away from the dichotomous thinking by which "an effect is statistically significant or it is not; it exists or it does not" (Cumming, 2014, p. 11).

Data are certainly informative beyond what NHST procedures look for. Reporting significant differences between means or significant correlations between variables reveals that the data display large departures from what one would generally observe if the null were true, but statistical significance does not imply theoretical or practical relevance. The latter is better judged by comparison with arbitrary but reasonable criteria on the magnitude of the effect, or via inference involving a plausible range for the parameter of interest. This is where ESs and CIs come into play, but they serve a purpose other than testing hypotheses and cannot be considered universal substitutes for it (Morey, Rouder, Verhagen, & Wagenmakers, 2014).

For an illustrative example, consider the search for behavioral correlates of the 2D:4D ratio. Many studies have reported significant (sometimes highly significant) correlations with aggression (see Hönekopp & Watson, 2011) and other behavioral measures in males and/or females (e.g., Del Giudice & Angeleri, 2016; Fink, Manning, & Neave, 2004; Manning & Fink, 2008; Millet & Dewitte, 2006; Ranson, Stratton, & Taylor, 2015; Richards, Stewart-Williams, & Reed, 2015), which again means that sample correlations were larger than one would expect in a sample from a population in which the variables are unrelated. Looking at the magnitude of the correlations (when these were reported) reveals that those that were significant were sometimes as low as .08 and hardly ever larger than .3, with very little (if any) theoretical value or practical utility. Rejecting the null hypothesis of absence of correlation does not require a large sample correlation, particularly when sample size is large (see Fig. 1a), but there is more to a meaningful relation than just bumping into a correlation that is significantly different from zero, particularly on consideration of the many factors that affect correlation statistics as measures of the relation between variables (García-Pérez & Núñez-Antón, 2013; Pernet, Wilcox, & Rousselet, 2013). This example does not reveal a problem of NHST but, rather, an improper use based on testing an irrelevant and uninformative null hypothesis coupled with the false notion that the magnitude of a correlation is less important than its significance in a test of the nil null.

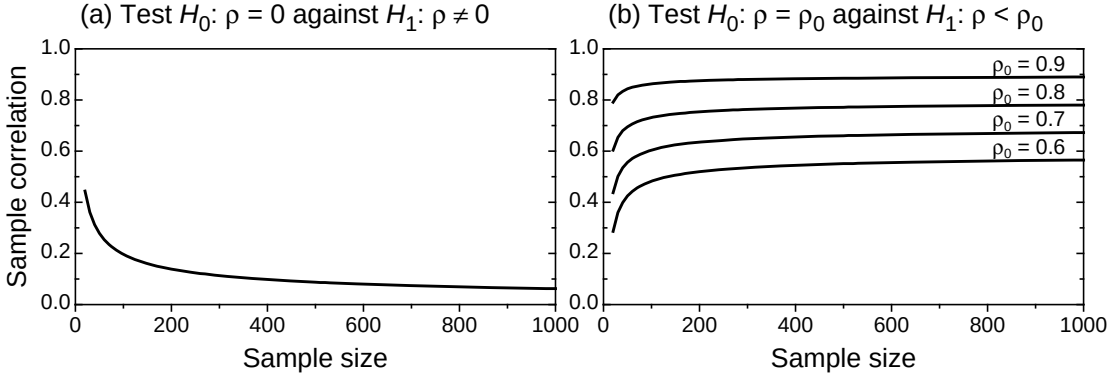

Figure 1. (a) Smallest value of the (unsigned) sample correlation that will reject the nil null hypothesis in a two-sided test with $\alpha = .05$, as a function of sample size. (b) Smallest value of the sample correlation that will not reject the non-nil null hypothesis in a left-tailed test with $\alpha = .05$, as a function of sample size and parameterized by the magnitude ρ_0 of correlation tested for (see the labels next to each curve). Sample correlations and criterion values ρ_0 are assumed positive, of course.

It certainly makes more sense to consider correlations in a criterion-referenced framework with a pre-defined lower bound on the magnitude that is regarded as empirically meaningful, subsequently judging the sample correlation obtained in a study from such perspective. Cumming (2014) makes a good point in this respect in defense of CIs for correlations (or differences between correlations) and also away from arbitrary and ill-defined cut points for their ESs. The same is true for proportions or differences between proportions, although CIs for proportions are very often computed incorrectly (García-Pérez, 2005). Standard deviations provide a reference to judge differences between means, but other statistics do not have such companions. For this reason, ESs are only well defined for means and their differences, where the meaningfulness of a difference is judged on consideration of the overlap between the distributions whose means are compared.

ESs for all other sample statistics have been defined arbitrarily and seemingly only with the purpose of filling up a look-up table of ESs for all conceivable statistics; the same holds for Cumming's (2014, p. 14) unelaborated suggestion that a measure of goodness of fit is the ES when the issue under investigation is how well a model fits the data. The resultant labels for ESs (e.g., small, medium, or large) are deceptive, if not plainly misleading. Computing and reporting ESs is thus informative only when the research questions imply means (or, more generally, location parameters): The ES naturally indicates how large the difference between means is, measured in units of standard deviation; in other cases, ESs do not have an actual referent and reporting them only comes down to reporting sample statistics accompanied by a label. But even in the cases in which reporting ESs makes sense, they are only point estimates based on the current data, with little value in the absence of an inferential framework that takes into account their sampling variability.

On their side, CIs are indeed calculated within an inferential framework that happens to be exactly that of NHST. The CIs that are typically calculated describe the range of parameter values that one might have placed in a null hypothesis that the current data would not have rejected in a two-sided test. Unpacking this statement for the case of differences between means, consider the non-nil null hypothesis H_0: $\mu_1 - \mu_2 = \mu_d$ tested against the alternative H_1: $\mu_1 - \mu_2 \neq \mu_d$ (compared to the conventional case in which μ_d is written out as 0). The CI for the difference between means is only the set of μ_ds for which the current data would not have rejected such null at the chosen significance level α. Every CI relates to null hypotheses on the corresponding parameter (or function of parameters) and carries only this meaning, although CIs thus defined have properties that allow other interpretations. CIs are thus NHST in a package. In all fairness to CIs, I must stress that they carry information that goes beyond what one gets from a simple two-sided test of the nil null hypothesis. For one thing, they make explicit that the data are compatible with many non-nil null hypotheses, not just the one that the researcher might have started off with.

Nevertheless, routine calculation of CIs is not guaranteed to address research questions adequately. Consider the example given by Cumming (2014, p. 20) upon discussing correlations. When the correlation of concern is an estimate of the reliability of a psychometric test, one may consider the test dependable only if its *true* reliability is at least .8. Realizing that true reliability can never be empirically estimated with the necessary precision, one would in these circumstances test the null hypothesis H_0: $\rho = \rho_0$ against the threat of a smaller value via the alternative H_1: $\rho < \rho_0$, with $\rho_0 = .8$. One would also hope that the data do not reject the null, for otherwise the psychometric test would rather be discarded or improved. Figure 1b shows the smallest value that the sample correlation might attain for not rejecting the null in this type of test, as a function of sample size and for selected ρ_0. (Most statistics textbooks describe how to perform this test, but no major statistical package seems to implement it; they only carry out a two-sided test of H_0: $\rho = 0$ without providing any CI.) If sample size was $n = 200$ and the reliability estimate was .75, the p-value is .039 and the hypothesis that $\rho = .8$ is rejected at the .05 significance level. The inappropriate (under the circumstances) 95% CI linked to a two-sided test spans the range from .68 to .81, which instead includes the target value of .8. A suitable one-sided CI would be needed in this case and surely in many others (e.g., when research questions involve mean differences in a specific direction; see Kaji & Lewis, 2010). This example should not be misconstrued as indicating that CIs are useless, it only stresses that routine calculation of two-sided CIs is bound to replicate the problems that improper use and misinterpretation of NHST has created. This example also shows that testing hypotheses is the natural approach to some research questions, which cannot be replaced with off-the-shelf alternative approaches.

Bayes factors

Wagenmakers (2007) dismissed NHST and its p-values and advocated a switch towards Bayesian inference and the quantification of statistical evidence via Bayes factors (see also Goodman, 1999a, 1999b, 2005). Subsequent papers developed these ideas and provided tools for the computation of Bayes factors in a few of

the cases encountered in psychological research (e.g., Morey & Rouder, 2011; Mulder, 2016; Nathoo & Masson, 2016; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, & Morey, 2009; Wetzels & Wagenmakers, 2012).

In a nutshell, Bayes factors are odds reflecting how much more likely the observed data are under the null hypothesis than under the alternative, or vice versa. (When the alternative is not a point hypothesis, a prior distribution on the parameter of concern must be assumed to carry out the computation so that it is not only the data that are contributing to the final outcome, although this is the least of the problems with Bayes factors.) If a priori probabilities for each hypothesis are additionally defined, the Bayes factor can be transformed into odds reflecting how much more likely one of the hypotheses is over the other, thus coming close to some scholars' ideal of attaching probabilities to hypotheses. This extra step also requires explicit assumptions about the a priori probability of each hypothesis. A common option is to assume equiprobability so that the Bayes factor is also the odds of the hypotheses. Bayesian inference via Bayes factors comes with no pre-defined cut points for dichotomous decisions, although rules of thumb have been defined to label them according to their magnitude, interpreted as strength of evidence (see table 3 in Wagenmakers, 2007, and table 1 in Wetzels et al., 2011). In the end, dichotomous decisions to reject or not to reject the null are not made. Instead, the outcomes are expressed as "the data are *x* times more likely under the null (or alternative) than under the alternative (or null)," as "the null (or alternative) is *x* times more likely than the alternative (or null) given the data," or as "the data display weak (or positive, or strong, or very strong) evidence in favor of the null (or alternative) hypothesis."

Undoubtedly, Bayes factors offer information that is not directly available under NHST, which is understandable on the basis that Bayes factors arise from a different framework. The Bayes factor as a measure of strength of evidence is a welcome addition to the researchers' toolkit, but it is another sample statistic that only describes what the current data say indirectly about the hypotheses. For the purpose of inference, one would need to know how likely it is that the Bayes factor indicates evidence against the null when the null is actually true, or how likely it is that the Bayes factor indicates evidence in favor of the null when the null is false. The probabilities of false positives and false negatives need to be known even though the Bayesian approach intentionally avoids dichotomous decisions. Remarkably, this aspect of Bayesian inference has been systematically overlooked (but see Gu, Hoijtink, & Mulder, 2016; Sanborn & Hills; 2014). Sample simulation results to this effect are presented next that systematize and extend some intriguing results reported by Wetzels et al. (2011), namely, that Bayes factors and *p*-values from 855 empirical studies are extremely tightly related, much more so than Bayes factors and effect sizes (see their figures 2 and 3).

One of these simulations was carried out in the context of paired-samples *t* tests (or, equivalently, one-sample *t* tests; code is available as Supplementary Material). One of the two factors in the design was sample size (n = 20, 40, 80, or 160); the second factor was the mean (μ = 0 or 5) of the normal distribution with standard deviation σ = 10 from which the data were drawn. Ten-thousand samples (replications) were drawn in each condition, and two outcome measures were obtained from each sample. One was the *p*-value for a test of H_0: $\mu = 0$ against H_1: $\mu \neq 0$; the other was the Bayes factor proposed for use in this situation by Rouder et al. (2009) in their equation 1. Note that the null is true when data are drawn from a distribution with mean 0 and that it is false (with an ES of 0.5) when data are drawn from a distribution with mean 5.

Figure 2 shows scatter plots of log Bayes factor against log *p*-value in each condition. Remarkably, Bayes factors and *p*-values have a one-to-one relation, and the same whether the null is true (open circles) or false (red crosses). The relation varies slightly with sample size: Note that the strings of symbols cross the vertical and horizontal dashed lines at their intersection in the leftmost panel (for n = 20), whereas that intersection is below the path of the symbols in the rightmost panel (for n = 160). Nevertheless, since n is a fixed value in any given study, the relation is perfectly determined; the slight misalignment that can be observed in figure 3 of Wetzels et al. (2011) is thus understandable as a result of the different sample sizes across studies and,

possibly, also due to round-off errors in the *p*-values and *t* statistics reported in the original sources. This Bayes factor does not carry any information that is not also in the *p*-value for given *n*, although it certainly makes it explicit in a way that is more readily interpretable. In other words, the Bayes factor is only a transformation of the *p*-value, something that can be anticipated from the fact that, like the *p*-value, the Bayes factor is determined by the value of the *t* statistic and the size *n* of the sample. Because of the one-to-one relation with *p*-values, labels such as "substantial," "strong," "very strong," or "decisive" evidence against the null (defined for Bayes factors) are only substitutes for analogous labels that might be defined for *p*-values (e.g., "barely significant," "significant," "highly significant," etc). Incidentally, Wagenmakers' (2007, p. 780) dismissed the probability of replication (p_{rep}; Killeen, 2005a, 2005b, 2006) on the basis that, by being a transformation of the *p*-value, "p_{rep} inherits all of the *p* value problems that are discussed in this article." If one accepts this argument, there is no option but to conclude that Bayes factors also inherit those problems and, hence, that they should be identically dismissed. But this simulation had a different purpose and other aspects of these results are worth discussing.

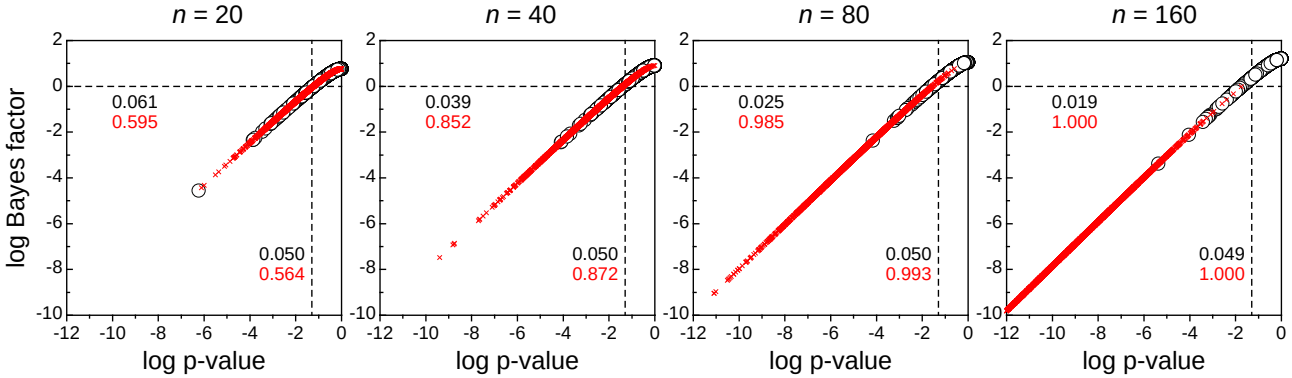

Figure 2. Scatter plots of log Bayes factor against log *p*-value for true (open circles) and false (red crosses) null hypotheses at four different sample sizes (panels) in a paired-samples (or one-sample) test for the mean.

In each panel of Fig. 2, the vertical dashed line indicates the conventional $\alpha = .05$ that sets the boundary for rejecting or not rejecting the null; the inset numerals near the bottom end of the line give the proportion of rejections according to this criterion when the null is true (black) and when the null is false (red). The black numerals thus reflect the empirical Type-I error rate, which stays at the stated level for all sample sizes; the red numerals instead reflect empirical power, which increases with sample size also as expected given an ES of 0.5 and $\alpha = .05$. Analogously, the horizontal dashed line in each panel indicates the Bayes factor of 1 that sets the boundary between evidence in favor or against the null; the inset numerals near the left end of the line give the proportion of times that evidence of any strength is found against the null. The black numerals thus indicate the false-positive rate (finding evidence against the null when the null is true), which decreases with sample size; the red numerals indicate the hit rate (finding evidence against the null when the null is false), which increases with sample size. It is remarkable that the hit rate of the Bayes factor is analogous to the power of the *t* test: The Bayes factor will indicate evidence in favor of a false null about as often as an NHST procedure on the same data will not reject the false null. Erroneous inferences occur in both cases, but NHST admits them and calibrates their (adjustable) rates whereas Bayes factors give instead the misleading outlook of solid and incontestable conclusions extracted from the data.

Another overlooked shortcoming of Bayes factors is apparent in the analysis of an additional simulation condition in which data were drawn identically but now from a distribution with mean $\mu = 2$, so that the ES is now small (i.e., 0.2). These results are not shown graphically but a one-to-one relation with *p*-values was naturally obtained again that fell on the same curves shown in Fig. 2. The distribution of Bayes factors was analyzed differently here. Specifically, I computed the proportion of samples in which the Bayes factor was

at or below 0.1, which would be interpreted as (at least) substantial evidence against the null (see table 1 in Wetzels et al., 2011, but keep in mind that their table reflects Bayes factors computed as the inverses of those used here). As sample size increased from 40 to 320 in multiplicative steps of 2, Bayes factors of 0.1 or smaller were obtained in 3.5%, 9.0%, 24.0%, and 60.5% of the samples. Thus, as sample size increases, evidence against the null becomes progressively more prevalent (and also stronger), regardless of how small the ES may be. A researcher who is content with reporting substantial (or more) evidence against the null overlooks that such evidence may only inform on a meaningless or negligible departure from the null. Because of the one-to-one correspondence with *p*-values, claims of substantial evidence against the null based on Bayes factors say no more than claims of large significance based on *p*-values. Both types of claim require additional consideration of the empirical meaningfulness of such departure from the null.

Bayes factors can incorporate expectations and goals in the form of a scale factor *r* for the prior on effect size (see Morey & Rouder, 2011; Rouder et al., 2009). This seems to detach Bayes factors from *p*-values, whose computation cannot be amended analogously. However, this scale parameter is only another element in the computation (see the amended equation in endnote 4 of Rouder et al., 2009) and, hence, Bayes factors and *p*-values still have a one-to-one relation similar to that shown in Fig. 2 above (where $r = 1$). The relation varies slightly across (reasonable) values for *r* but, again, given the researcher's choice on it, the relation with *p*-values is perfectly determined. In the absence of an analysis of the resultant error rates, it is unclear whether using a non-unit value for parameter *r* is advisable, but such analysis will not be pursued here.

These characteristics were confirmed in an analogous simulation involving correlations. One of the factors in this design was sample size ($n = 20$, 40, 80, or 160); the second factor was the correlation ($\rho = 0$ or .3) between variables whose distribution was bivariate normal. Ten-thousand samples were again drawn in each condition and analogous outcome measures were obtained from each sample: The *p*-value for a test of H_0: $\rho = 0$ against H_1: $\rho \neq 0$ and the Bayes factor proposed by Wetzels and Wagenmakers (2012) in their equation 13, although it was computed here to reflect evidence for the null. Note that the null is true for data drawn from a distribution with $\rho = 0$ and that it is false for data drawn from a distribution with $\rho = .3$. Scatter plots of log Bayes factor against log *p*-value in Fig. 3 reveal the same pattern shown in Fig. 2: A one-to-one correspondence between Bayes factors and *p*-values (understandable again because both of them are determined only by the sample value of the correlation and the size of the sample) and hit rates for the Bayes factor that are comparable to the power of the *t* test to detect non-null correlation under NHST.

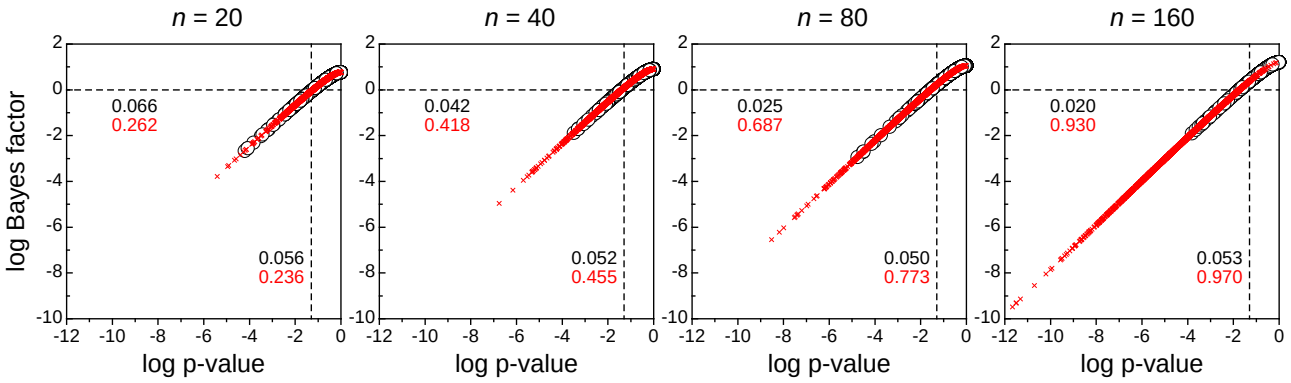

Figure 3. Scatter plots of log Bayes factor against log *p*-value for true (open circles) and false (red crosses) null hypotheses at four different sample sizes (panels) in a test for the correlation between two variables.

What these simulations illustrate is that Bayes factors and *p*-values are interchangeable. Both arise from testing against the data two mutually exclusive hypotheses about the underlying reality from which the data were drawn, and they only express the outcome in different terms. The Bayes factor (with or without extraneous information coming from a prior on effect size) is only another sample statistic with some

sampling distribution and, thus, with a non-negligible variability across samples. The current sample of data can provide Bayesian evidence in favor or against the null that opposes the true reality of the null. Moreover, when the null is false but the true ES is very small and sample size is large, Bayes factors will indicate decisive evidence against the null while entirely missing that the departure from the null is empirically meaningless (see figure 5 in Rouder et al., 2009). This outcome is not different from the extremely small *p*-values that would have been obtained via NHST procedures in these cases. Then, Bayes factors cannot be taken at face value and the intuitive appeal of their wording as strength of evidence or odds of hypotheses is misleading, unless researchers are conscientious of the lack of an inferential framework and phrase their conclusion in a form similar to "well, this is what came out of our sample; who knows what would have happened if we had come across a different sample from the same population?" A suitable inferential framework for Bayes factors is still needed because researchers must report the error rates of their polytomous claims of evidence in favor or against the null (García-Pérez, 2012).

Ban on NHST

In a bold move, the journal *Basic and Applied Social Psychology* (BASP) recently banned NHST and CIs (Trafimow & Marks, 2015). Manuscripts submitted to the journal can include NHST results at the authors' discretion, but all of them should be removed later so that no references to test statistics, *p*-values, or statistical significance appear in published papers. The ban is founded on the claim that NHST is invalid and cannot provide what allegedly matters to researchers, namely, the probability that the null hypothesis is false. CIs are also banned because their interpretation relies on an inverse logic that is tagged as irrelevant to researchers' goals. Bayesian approaches to data analysis are not banned but the journal reserves the right to judge their appropriateness case by case. Quoting literally from Trafimow and Marks, only "strong descriptive statistics, including effect sizes" are permitted, and the ban is expected to have the effect of "increasing the quality of submitted manuscripts." A subsequent paper (Valentine, Aloe, & Lau, 2015) gives guidance as to how to analyze data and present results under the new policy, and a follow-up editorial (Trafimow & Marks, 2016) further develops the arguments behind the ban.

One can only sympathize with an initiative that favors close scrutiny of the data to highlight distributional aspects via descriptive statistics, instead of reporting those statistics (if at all) as mere companions to NHST analyses and simple claims of significance or lack thereof. Considering sampling variability, descriptive statistics are unlikely to reflect other than the characteristics of the current samples unless those samples are very large and representative. In fact, the new policy encourages the use of large samples. Large samples (e.g., from internet studies) are available in some areas of research and, in those cases, descriptive statistics can arguably be taken at face value. What is nevertheless unclear under the new editorial policy is how one would go from sample statistics to inferences of any type in cases in which samples are still small, how the probability that the null hypothesis is false would be determined (which is what allegedly matters), or how the conclusions of the study would be worded in a way that avoids the infamous dichotomous thinking. In principle, because formal and principled methods of inference are not used, reported results can only be taken as a simple and unpretentious description of the current sample of data.

To search for clues about how inference is handled in these conditions, I looked at the 44 papers published thus far in BASP under the new policy (Volume 37, Issues 1–6, 2015, Volume 38, Issues 1–2, 2016, and three unpublished online papers). Some of those papers were editorials, introductions to special issues, theoretical or methodological papers, meta-analyses, reviews, or commentaries; only 27 papers reported research results. The analysis of reporting practices in those 27 papers was very disappointing: Stripping off NHST does not seem to have swept along the way the type of inferences and claims that are commonplace under the conventional misuse of NHST and misconstruction of its outcomes. Factual categorical conclusions appear in the title of 9 of the 27 papers (e.g., "*Self-control depletion does not*

diminish attitudes about being prosocial but does diminish prosocial behaviors"). Factual categorical conclusions are also found in the abstract of 24 of the 27 papers (e.g., "*implicit attitude transfer … is not disrupted by intentional control*"). Finally, factual categorical conclusions are found in all papers after almost each of the individual analyses for which a descriptive statistic or ES is reported. Quite regrettably, the only noticeable difference that the new policy has brought with respect to papers published elsewhere (or earlier in BASP) is the absence of NHST results or jargon; "inferences" now come straight out of descriptive statistics and ESs interpreted at face value without a proper inferential framework. I have not identified in the papers any sign that use of NHST would have led to different statements or conclusions, or that it would have made those statements wrong or inappropriate given the data. The new policy seems business as usual without the burden of NHST. I concede that it is perhaps too early to judge how the new policy allows determining the probability that hypotheses are true or false and brings research of higher quality or conclusions that are more dependable, meaningful, or "correct," but the dichotomous thinking lives on and seems welcome.

It is also important to stress that a dogmatic ban on NHST has overly undesirable consequences. In two remarkable cases among those papers, research questions that called for tests of homogeneity of distributions were instead addressed by reporting thoroughly inappropriate correlations or isolated odds ratios, the only descriptive statistics that sound like putting together information from two variables. Although they were posited in other terms, the research questions were whether or not some variable has the same distribution in two groups, which is unrelated to the questions of whether that variable correlates with group membership or whether sample odds ratios differ here or there. Since no sample statistic (let alone ES) can inform of homogeneity of distributions in an interpretable manner, a policy that only permits descriptive statistics and ESs forces researchers to use inadequate methods that cannot logically answer their research questions. Yet, the questions are regarded as answered and the answers (to other questions) are interpreted in the context of the actual research questions, apparently to the satisfaction of BASP editors and reviewers.

Summary of alternatives to NHST

The advantages of the three alternatives just discussed were presented by their proponents in sharp contrast with selected features of NHST that were stressed as problematic. However, the virtues of the alternative methods were usually presented without consideration that the data collected in any given study are only a random sample from a population of observations. Thus, any summary measure defined and computed as some function of the data (sample statistics, *p*-values, CIs, ESs, Bayes factors, etc) is a random variable whose sampling distribution is of the utmost importance in any attempt at interpreting it or at making inferences. Across replications, data are newly drawn independently from the population and new values for the summary measures are obtained which come from their corresponding distributions. Replications thus reflect the distribution of the summary measures and they can only be properly judged in that context. An example will illustrate.

As discussed earlier, the proportion of significant *p*-values (i.e., those that are lower than α) should be close to α across replications (for any and all possible α) if the null hypothesis is true, whereas it should be close to the computed power $1 - \beta$ if the null is false given some ES. Thus, the subjective belief in the truth of the null is mediated by the observed proportion of significant *p*-values reported across replications (provided there is no publication bias, of course). Consider the illustration in figure 1 of Cumming (2014). Twenty-five replications were simulated of an experiment involving two groups of size 32 each with normally-distributed observations in each population, an ES of 0.5, and power of .5 at $\alpha = .05$. The null hypothesis of equality of means was rejected by the criterion of a significant *p*-value in 12 of the 25 replications, which is as close as one can get to the expected proportion of rejections in these conditions (i.e., the power of .5). Thus, replications provide the evidence needed to judge the (un)tenability of the null and to

estimate the ES that would have produced such proportion of rejections (power) at the selected α.

This result is in sharp contrast with the analysis of the same *p*-values reported by Cumming (2014), which was founded on the notion that a replication is successful if it renders the same decision as the study it replicates. In his words, the principle guiding his analysis was that "if *p* reveals truth ... then replication *p* should presumably reveal the same truth" (Cumming, 2014, p. 12), a proposition based on the incorrect notion that *p*-values reveal truth and, thus, that they should be consistently on the same side of the boundary across replications. In any case, by this criterion only 9 of the 24 successive pairs (38%) represented successful replications. In contrast, an analysis of the corresponding CIs, for which success was now defined as the mean difference obtained in a replication falling within the CI of the original study, rendered a success rate of 83.3% (20 of the 24 successive pairs). I am still to be convinced that Cumming's analysis reveals any problem with *p*-values, that it provides any evidence that CIs reveal truth (which *p*-values certainly do not), or that it shows CIs to be better inferential tools than *p*-values, unless one accepts that the goal of replications is only to confirm the decision made in the original study and, thus, that a method of analysis should be used with which the probability of tagging a replication as successful is large. Interestingly, 23 of the 25 CIs (92%) contained the true mean difference, which is again close to the expected 95% coverage. The trouble with CIs in practical analyses of replication studies is that one does not have the information needed for this assessment (i.e., the true value of the population parameter), nor does one have any means to figure out which of the CIs were those that include the true value of the parameter so as to narrow down the range where the true parameter lies.

A second characteristic of the alternatives to NHST is that they focus on a relatively narrow set of situations, namely, when interest lies in distributional parameters (mostly means, proportions, correlations, and their differences). ESs or CIs are not the ultimate solution, but they are valuable only *when inference about parameters is involved*. The NHST framework is broader and includes, for instance, procedures for testing hypotheses about the form of distributions, or procedures for the analysis of contingency tables that permit tests of independence (in various forms) and tests of homogeneity of distributions. When these are the issues under investigation, inference about parameters is not involved and ESs and CIs cannot be defined. An even broader area in which NHST cannot be replaced with ESs, CIs, or other alternatives is the assessment of the goodness of fit of process models, an issue discussed in the next section.

Assessing goodness of fit

The debate over the utility and adequacy of NHST and the search for alternative methods of analysis has focused on inference about distributional parameters, systematically overlooking the growing approach to research that consists of testing quantitative models of psychological processes (Busemeyer & Diederich, 2010; Cavagnaro, Myung, & Pitt, 2013; Estes, 1993; Farrell & Lewandowsky, 2015; Luce, 1985, 1995). In this approach, outcome variables are viewed as the observable manifestation of unobservable processes whose functional characterization is the actual object of study. Quantitative models are thus set up that express an observable outcome variable as the final result of the integrated operation of distinct unobservable processes characterized by interpretable latent parameters. Very few (if any) of the parameters of the distribution of the outcome variable itself are relevant or subjected to inferential analyses under this approach. Instead, the functional description embodied in the model is the hypothesis that is tested in an attempt to determine its empirical validity. The term "empirical validity" refers to the capability of the model to account for observed data, suggesting neither that the model is true nor that no alternative model could also account for the data.

Testing the validity of such models requires assessing the extent to which they fit a given data set with suitably estimated latent parameters for the hypothetical processes. Since model parameters surely vary across individuals, the model under scrutiny is fitted so as to estimate separate parameter values for each

participant. This statement should not be misconstrued as a denial that this approach is sometimes placed in a broader hierarchical or multilevel context with group-level parameters (e.g., to model individual differences; Bartlema, Lee, Wetzels, & Vanpaemel, 2014) simultaneously estimated with the individual-level parameters (for the process model) implied in this discussion. This statement should also not be misconstrued as a denial that sometimes data are aggregated across individuals before fitting the model to average performance, a strategy that may be useful for some purposes but that is inappropriate for testing the validity of a process model on account of individual differences in the hypothesized processes (see, e.g., Estes, 1956; Estes & Maddox, 2005). With non-aggregated data, each individual represents a replication in a study aimed at checking out the tenability of the model. Hence, under NHST, the collection of *p*-values for goodness-of-fit statistic across participants provides the information needed to assess empirical validity. If the goodness-of-fit test rejects the model (at some α) in about 100α% of the individual cases, the model gains support and a subsequent analysis of parameter estimates is warranted; if the test instead rejects the model in a large percentage of cases, the model ought to be discarded or amended with no additional analyses of the estimated parameters. In such cases, appropriate residual analyses provide useful indication of where and how the model might need to be amended (García-Pérez, Núñez-Antón, & Alcalá-Quintana, 2015).

Numerous examples in the literature illustrate this approach to psychological research in a diversity of areas, including diffusion models for decision and reaction times (Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Smith, 2015; Smith & Ratcliff, 2015; Voss, Nagler, & Lerche, 2013), multinomial process tree models for cognitive processes (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Riefer & Batchelder, 1988), or decision models of perceptual judgments (García-Pérez & Alcalá-Quintana, 2011, 2012; Sridharan, Steinmetz, Moore, & Knudsen, 2014; Yang, Meijer, Doll, Buitenweg, & van Gils, 2015). Provided that the model fits the data, the researcher looks at parameter estimates to identify the particular process(es) that may produce observed differences in the outcome variable between groups (e.g., Chechile, 2010; Pe, Vandekerckhove, & Kuppens, 2013; Ratcliff & van Dongen, 2009; Ratcliff, Perea, Colangelo, & Buchanan, 2004; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002; van Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmakers, 2014). Moreover, when within-subjects conditions are included in the study that presumably affect only a subset of the processes, the model can be fitted in search for an account of the data in which parameters relating to processes that are not expected to vary across conditions are estimated to have the same values (though still subject to individual differences) whereas parameters that presumably vary across conditions are separately estimated for each condition (see, e.g., Allik, Toom, & Rauk, 2014; García-Pérez & Alcalá-Quintana, 2012, 2015a, 2015b, 2015c; García-Pérez & Peli, 2014, 2015; Mueller & Kuchinke, 2016). In these circumstances, if the goodness-of-fit statistic still rejects the model approximately the stated number of times at the chosen α, the empirical validity of the model gains further support.

Because these models include parameters that relate to the various processes determining observed performance (and once the model has been shown to fit the data adequately), model parameters estimated for each participant and condition allow inferences about the processes that each parameter represents, and also about how those processes vary across groups or across conditions. Such subsequent analyses obviously call for NHST procedures or their alternatives and, from this point on, inference goes as discussed earlier. Nevertheless, the first stage assessing the extent to which the model fits the data can only be carried out (so far) via NHST procedures or alternatives that are essentially identical. For instance, sometimes the situation on hand is unlikely to meet the conditions justifying the use to the asymptotic distribution of goodness-of-fit statistics. In such cases, *p*-values are instead obtained via *parametric bootstrap*, which consists of generating a large number of simulated data sets from the model using the parameter values that were estimated from the empirical data set. The collection of values for the goodness-of-fit statistic across all simulated data sets thus gives the sampling distribution of the test statistic, from which the *p*-value for the empirical data is obtained (see, e.g., Wichmann & Hill, 2001). Similarly, if Bayesian estimation was used, *predictive checks*

imply an analogous strategy in which data are replicated using the predictive distribution with parameter values drawn from the posterior distribution, which provide the context for the computation of a predictive *p*-value (Meng, 1994; for a slightly different approach, see Batchelder & Anders, 2012). To the best of my knowledge, predictive *p*-values (or other principled approach with known error rates) have not been used to assess the goodness of fit of process models: Predictive checks generally consist of displaying data and model predictions for a visual judgment of agreement (e.g., Bartlema et al., 2014). On the other hand, the use of predictive *p*-values in the context of inference about distributional parameters is not without problems (see Ames, 2015; Gelman, 2013; Lee, 2011).

A relatively side issue in model-based analyses arises when alternative models compete to account for the same data. Models can never be proven true or false: They are only compatible or incompatible with empirical data and, of course, alternative models that hypothesize different underlying mechanisms, processes, or operations can be (statistically) equally compatible with the data. In principle, the comparative validity of competing models could always be assessed in a crucial experiment in which the models make different predictions so that data thus collected will allow ruling out one of them. However, such crucial experiments cannot always be identified (or cannot be carried out) and, hence, an alternative basis for model comparisons is needed when the incumbent models fit the data sufficiently well in the first place. This is another area in which fully-developed alternatives to NHST (i.e., under a proper inferential framework with control of error rates) do not exist, and where the (proper) use of NHST methods is perfectly justifiable.

A popular non-NHST approach to model comparisons consists of using metrics that combine additively some measure of misfit and some penalty based on the complexity of the model (see, e.g., Turner, Sederberg, & McClelland, 2016; Vandekerckhove, Matzke, & Wagenmakers, 2015). The measure of misfit increases with poorness of fit and the complexity penalty increases with the number of parameters in the model. This practice sets an economical criterion in which parsimony (i.e., mere simplicity of description) is given the same weight as the extent to which the model is capable of giving an acceptable account of the data. Subsequently, a dichotomous decision is made and the model whose sample measure on this metric is smallest gets selected. Because these binary decisions are usually made without consideration of the quality of the fit on its own, it is not unlikely that a poor-fitting simple model with few parameters outperforms a good-fitting and more complex model that uses more parameters. The binary decision is based on an absolute comparison of sample statistics for each model and is, thus, entirely detached from a proper inferential framework in which the sampling distributions of these statistics are taken into account. In a sense, using the chosen metric to select models in this way is analogous to concluding that the population with a higher mean on some variable is that whose sample of individuals in a study had a higher mean.

For an illustration that follows up on an observation to this effect reported by García-Pérez and Alcalá-Quintana (2012), consider the case of models of performance in temporal-order judgment tasks, in which two stimuli (A and B) are presented in each trial with some stimulus onset asynchrony (SOA) and observers must report whether A or B was presented first. A two-parameter model might posit that the proportion *P* of "A first" responses as a function of SOA (Δt) is given by

$$P(\Delta t) = \frac{1}{1+\exp[-\beta(\Delta t - \theta)]}, \tag{1}$$

where θ and β are the two model parameters. A more elaborate process model (for details, see García-Pérez & Alcalá-Quintana, 2012) posits that the relation to SOA is instead given by

$$P(\Delta t) = 1 - (1 - \xi)F(\delta; \Delta t) - F(-\delta; \Delta t), \tag{2}$$

where

$$F(d;\Delta t)=\begin{cases}\dfrac{\lambda_A}{\lambda_A+\lambda_B}\exp[\lambda_A(d-\Delta t-\tau)] & \text{if } d\le\Delta t+\tau\\ 1-\dfrac{\lambda_B}{\lambda_A+\lambda_B}\exp[-\lambda_B(d-\Delta t-\tau)] & \text{if } d>\Delta t+\tau\end{cases} \tag{3}$$

and λ_A, λ_B, τ, δ, and ξ are the five model parameters. To assess the outcomes of this model-comparison approach, a simulation was conducted involving 10,000 replicates. Data for each replicate were generated with the model of Equation 2. In each replicate, true parameter values were randomly drawn from independent uniform distributions over [1/60, 1/40] for λ_A and λ_B, over [−20, 20] for τ, over [80, 150] for δ, and over [.3, .7] for ξ. Empirical proportions of "A first" responses were then simulated for each replicate along 40 trials at each of 15 SOAs (from −280 ms to 280 ms in steps of 40 ms; SOA is defined as positive when stimulus A is presented before stimulus B). The models of Equation 1 (a false model) and Equation 2 (the true model) were then separately fitted to data from each replicate via maximum-likelihood estimation of their parameters, subsequently computing in each case the likelihood-ratio statistic G^2 (an appropriate NHST measure of goodness of fit in these conditions) and the Bayesian information criterion (BIC; a measure of economy of description).

The results confirmed that model selection based on the principle of economy favors poor-fitting models that incur a smaller penalty for complexity. At $\alpha = .05$, the generating five-parameter model of Equation 2 was rejected by the G^2 statistic in 4.78% of the cases (virtually the expected rejection rate given that the model is true) whereas the simpler (and false) two-parameter model of Equation 1 was rejected instead in 61.2% of the cases. Despite this massive rejection rate, the poor-fitting two-parameter model outperformed the five-parameter model by the BIC in 19.95% of the cases. Figure 4 shows an illustrative example from one of the replicates. This simulation stresses two points. One is that NHST procedures are well suited to assessing goodness-of-fit and model comparisons via the empirical proportion of significant *p*-values across replicates, which are readily computed when models are fitted individually to each participant and condition within or across studies (see, e.g., García-Pérez & Alcalá-Quintana, 2015a, 2015b, 2015c). The second is that model-comparison approaches based on metrics detached from an inferential framework are bound to render incorrect dichotomous decisions that capitalize on sampling variability and do not do justice to the data.

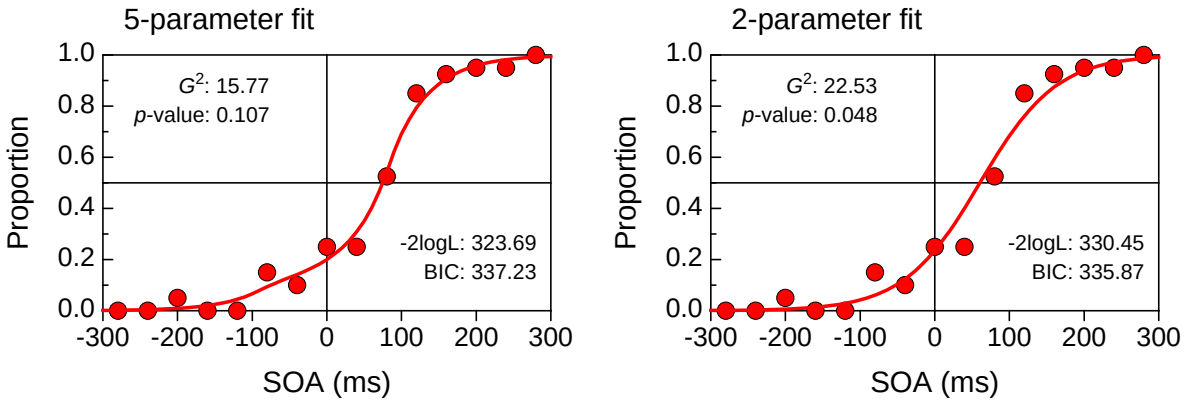

Figure 4. Sample data (symbols) and fitted curves from the models in Equation 2 (left panel) and Equation 1 (right panel). Insets show the value and *p*-value of the goodness-of-fit statistic G^2 (with 10 degrees of freedom in the left panel and with 13 degrees of freedom in the right panel), the measure of misfit given by twice the value of the negative log-likelihood of the data under each model ($-2\log L$), and the value of the BIC, which adds to the value of $-2\log L$ a penalty of ln(15) units per parameter in the model. The two-parameter model is rejected by the G^2 statistic but it nevertheless outperforms the five-parameter model by the BIC.

The preceding simulation illustrates the inadequacy of model selection based on the BIC without concern for a preliminary assessment of whether the data reject the models in the first place. In practice, the reality generating the data is unknown and it is conceivable that none of the models that a researcher considers is adequate (i.e., that they all will be rejected by the data far more often than expected for a true model). Oblivious to this, the BIC will nevertheless select one of those (incorrect) models, and researchers may even classify observers according to which model "holds" for them. The same inadequacy arises with other model-selection criteria that also overlook whether the data reject the models to be compared (e.g., the use of Bayes factors for this purpose; see, e.g., Kary, Taylor, & Donkin, 2016; Myung & Pitt, 1997).

Misuse of NHST

Although a discussion of the misuse of NHST was declared not to be the goal of this paper, there are a couple of related issues that are worth commenting on in the context of this paper. There is also a specific misuse of NHST that deserves commentary because it has not been discussed earlier and needs to be eradicated. Both issues are discussed next.

The nil null hypothesis and the point hypothesis

One of the most common objections to NHST is that it is a methodology designed to test nil null hypothesis (i.e., hypotheses in which the parameter value is zero), when researchers are rarely interested in them and the nil hypothesis is known to be false almost with certainty (Cohen, 1994; Schneider, 2015). Consideration of nil null hypotheses is certainly common in the statistics textbooks used in courses aimed at training today's students (some of whom will be tomorrow's researchers) and this is also the only type of hypothesis that the most popular statistical packages entertain. However, NHST procedures are by no means limited to such cases. A related objection is that NHST involves point null hypotheses, of which the nil null is a particular case. Any hypothesis involving a single value of a continuous parameter is false by definition: The probability that the population parameter attains that particular value is zero, and this holds for nil and non-nil values of the parameter. Objections to NHST raised on consideration of (nil or non-nil) point hypotheses thus focus on the limitations of statistical software packages and not on those of NHST itself.

Consider a researcher interested in investigating whether treatment 1 is superior to treatment 2 such that the average value of a relevant outcome measure under treatment 1 exceeds that under treatment 2 by some meaningful amount at the least. NHST permits testing the null H_0: $\mu_1 - \mu_2 = \mu_d$ against the alternative H_0: $\mu_1 - \mu_2 > \mu_d$, for any arbitrary $\mu_d > 0$ selected by the researcher on the basis of the issue under investigation. The distribution of the test statistic is well defined in this case too and the properties of the decision rule are not affected. Incidentally, in a situation like this, one might also follow the approach of Jones and Tukey (2000). The procedure will thus assess the statistical significance of a difference in means that is at least as large as the value of μ_d that is deemed relevant. But one cannot directly carry out this test using the most common, general-purpose statistical packages. Again, NHST is blamed for not allowing something that only widespread packages cannot do directly. The same holds in the assessment of relations between variables, where reliance on statistical software packages limits researchers to testing against absence of correlation.

This narrow focus misses consideration of NHST developments aimed at addressing interval hypotheses. These developments—which place NHST in a broader framework that considers hypotheses of equality, superiority, noninferiority, and equivalence—found home in the biomedical sciences (see, e.g., D'Agostino, 2002; Piaggio et al., 2006; Tunes da Silva, Logan, & Klein, 2009) and they are well covered in a number of reference sources some of which include computer code (e.g., Ng, 2014; Pardo, 2014; Rothmann, Wiens, & Chan, 2011; Wellek, 2010), although they have not yet made it into classical software packages for statistical analyses. Use of an inadequate, classical approach given the research question is not without consequences (see Kaji & Lewis, 2010) but a recent analysis (Allan & Cribbie, 2013) revealed that these developments and

the consequences of inadequate use of classical procedures have passed by researchers in psychology.

The foregoing discussing notwithstanding, questions such as “Is treatment 1 better than treatment 2?” should perhaps be regarded as marginally scientific: It is highly unlikely that some treatment is uniformly better than some other. Then, how one analyzes the data that would presumably answer such question is immaterial. Approaching this question from a perspective other than a search for mere differences in means (via NHST or its alternatives) is necessary to address comparisons of treatments with some chances of finding informative answers (see, e.g., Dusseldorp & Van Mechelen, 2014; Loh, He, & Man, 2015).

Use of NHST procedures to check for matched samples

Matched-samples designs are commonplace in psychological research when within-subjects studies are unfeasible, that is, when the study addresses differences between patients and normal controls, differences between alternative treatments or interventions, differences between groups defined according to subject variables, or differences in other situations in which control groups are used. Consideration of internal validity requires that groups are matched on extraneous variables that might contaminate the outcomes.

When random assignment to conditions is feasible, groups that are matched on extraneous variables can reasonably be created by first classifying participants on the basis of a measure of the extraneous variable and then randomly distributing the participants in each class evenly across the various groups. In the event that participants with the exact same values on the extraneous variable had been found, the groups would differ only by the error with which the extraneous variable has been measured. When random assignment is not feasible (e.g., when groups of patients and normal controls are involved), matching cannot be that precise and differences across groups may arise. The universal approach to checking for matched groups in these situations consists of using NHST procedures to test for equality of means across groups on each of the extraneous variables. Groups are then regarded as matched if the null is not rejected on any of these tests, with or without a correction for multiple testing. This practice is unjustifiable for two reasons.

Firstly, NHST procedures do not inform of differences between samples. They are instead set up to judge the tenability of differences at the population level given the observed differences between the samples. The hypothesis that is rejected or not rejected concerns population parameters and, thus, a non-significant result only says that the observed differences between the samples are compatible with the surmise that population parameters do not differ, which is irrelevant in this context. Secondly, when the concern is whether or not two (or more) particular samples differ on some variable, the samples exhaust the observations under consideration and (in NHST terms) they become the full populations under scrutiny: There is nowhere to go beyond the samples themselves. Since no inference is required in a direct comparison of populations, NHST procedures are not applicable. Judgments about the effective match of samples on extraneous variables should instead be based on the relevance or empirical meaningfulness of the observed differences. The criteria used in these judgments may come from other statistical analyses (e.g., the smallest difference that previous research has shown to be relevant, the largest inconsequential difference in test scores according to the reliability of the instrument, etc) but the question of whether samples are matched should never be (or have been) addressed using NHST procedures.

Playing the devil’s advocate, I might counter that use of NHST procedures for this purpose is only irrelevant. Yet, the practice is counterproductive by giving the false impression that samples are matched (if significant differences are not found) when they may not but, more importantly, by perpetuating uneducated misuse of NHST procedures and preventing conscientious consideration of relevant criteria to declare samples matched or unmatched.

Conclusion

When the characteristics of a population of observations are known, the problem of establishing what

characteristics samples might show is easily solved via direct deduction. Empirical research aims at driving this road in reverse, making inferences from whatever was observed in a single sample drawn from a population of unknown characteristics. Inference, including statistical inference, is an inverse problem that cannot be solved without error. NHST is one of the possible inferential approaches in these circumstances, and one that has undeniably been abused and misused by explicitly or implicitly presenting its outcomes as the final answers. But, as Cohen (1994) stressed, alternatives to NHST do not exist that would by magic allow error-free inferences. This does not seem to have deterred critics of NHST to advocate alternative approaches, although it is unclear whether they were proposed on the belief that the impossible may be achieved or only to instate alternative rites in hopes that misuse and misinterpretation do not also infect them. Nevertheless, passion has usually overridden objectivity in presentations of alternatives to NHST. Almost invariably, the charge against NHST has focused on misuse and misinterpretation with emphasis also on what NHST cannot deliver; ironically, this was followed by presentations of the proposed alternatives that neglected to mention that they cannot deliver either, most notably when the alternative is entirely detached from a proper inferential framework (e.g., Bayes factors or the BASP ban).

NHST has been purported to be the cause of all of our problems (e.g., the lack of progress in psychology, the replication crisis, the file-drawer effect, etc; see Branch, 2014), overlooking that data analysis is only the last stage of a process which can also go wrong in each of the preceding steps (Leek & Peng, 2015) and that alternative approaches are not immune to misuse or misinterpretation. Miller (1965, p. 18) claimed that "in science, at least half the battle is won when we start to ask the right questions." The rest is surely won when we look for the answers using dependable research methods to collect data (including careful experimental design with control of extraneous variables and confounds when required) and when we analyze the data in a sensible manner that addresses the posed research questions (whichever that manner happens to be in each case). The debate over NHST has been dominated by the conception that NHST is inherently misguided, incapable of providing any answer that we might care for, and urgently needing a replacement. Those points were made by discussing widespread misinterpretation of NHST and misconstruction of its results, which were often presented as problems of NHST itself. There is undeniable merit in some of the proposals that have been developed from alternative theoretical frameworks, which is one positive thing that the unfair charge against NHST has produced. Considering the threats that limit the validity of inference of any type, there is surely no universal approach to data analysis that will handle all research questions as each of them deserves. Abandoning NHST to embrace dogmatically an alternative approach does not guard against misuse and misconstruction, nor does it guarantee that research questions will be addressed adequately, that data will be analyzed as the research question demands, or that inference will be more dependable and revealing. What seems to be needed is not a replacement for NHST but the eradication of poor understanding of inference and of misuse or misinterpretation of its outcomes, whichever approach is used for inference.

Better practice is certainly needed and one might fantasize that researchers will make a move in that direction if so forced by the journals where they submit their manuscripts. Yet, journals are ultimately handled by researchers under a temporarily different role (as editors or reviewers) and adoption of these roles does not endow them with abilities that they do not display in their own research. Then, journals are part of the problem rather than vehicles for a solution. To illustrate with my most recent experience, I just witnessed an exchange in which a reviewer objected to a manuscript that reported a number of non-significant results (with large *p*-values) by arguing strenuously that the Type-I error rate was unacceptably large. That the *p*-value is *not* the probability of a Type-I error should be clear to anyone who understands NHST or who has at least digested authoritative commentary to this effect (see, e.g., Hubbard, 2011; Kline, 2004, Ch. 3), besides the fact that, by definition, a Type-I error has not occurred if the null was not rejected. Analogously, in a paper aimed to "promote clear thinking and clear writing (...) by curbing terminological misinformation and confusion," Lilienfeld et al. (2015) listed 50 terms and phrases that researchers should avoid. Their 22nd

item was the expression "$p = 0.000$," which they declared inappropriate because it "implies erroneously that there is a *zero* probability that the investigators have committed a Type I error" (original italics). Although writing out $p = 0.000$ is undoubtedly incorrect, the *p*-value is again not the probability of a Type-I error. The prevalence of this and other misconceptions despite numerous efforts at eradicating them over the years is dismaying. Without statistical education, promoting (or forcing) changes in the way data are analyzed and in the way results are reported will merely amount to embracing different rituals and worshipping different idols to continue making the same errors.

Acknowledgments

This research was supported by grants PSI2012-32903 and PSI2015-67162-P (Ministerio de Economía y Competitividad, Spain). Simulations were carried out on EOLO, the MECD- and MICINN-funded HPC of Climate Change at Moncloa Campus of International Excellence, Universidad Complutense.

References

Allan, T. A., Cribbie, R. A. (2013). Evaluating the equivalence of, or difference between, psychological treatments: An exploration of recent intervention studies. *Canadian Journal of Behavioural Science*, *45*, 320–328.

Allik, J., Toom, M., & Rauk, M. (2014). Detection and identification of spatial offset: Double-judgment psychophysics revisited. *Attention, Perception, & Psychophysics*, *76*, 2575–2583.

Ames, A. J. (2015). *Bayesian model criticism: Prior sensitivity of the posterior predictive checks method*. Doctoral Dissertation, University of North Carolina at Greensboro.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory and Psychology*, *24*, 256–277.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling*. Thousand Oaks, CA: SAGE.

Cavagnaro, D. R., Myung, J. I., & Pitt, M. A., (2013). Mathematical modeling. In T. D. Little (ed.), *The Oxford Handbook of Quantitative Methods. Volume 1: Foundations*, pp. 438–453. New York: Oxford University Press.

Chechile, R. A. (2010). Modeling storage and retrieval processes with clinical populations with applications examining alcohol-induced amnesia and Korsakoff amnesia. *Journal of Mathematical Psychology*, *54*, 150–166.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

D'Agostino, R. B. (2002). Non-inferiority trials: Advances in concepts and methodology. *Statistics in Medicine*, *22*, 165–167.

Del Giudice, M., & Angeleri, R. (2016). Digit ratio (2D:4D) and attachment styles in middle childhood: Indirect evidence for an organizational effect of sex hormones. *Adaptive Human Behavior and Physiology*, *2*, 1–10.

Dusseldorp, E., & Van Mechelen, I. (2014). Qualitative interaction trees: A tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine*, *33*, 219–237.

Erdfelder, E., Auer, T.-S., Hilbig, B.E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie*, *217*, 108–124.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

Estes, W. K. (1993). Mathematical models in psychology. In G. Keren & C. Lewis (eds), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, pp. 3–19. Mahwah, NJ: Erlbaum.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408.

Farrell, S., & Lewandowsky, S. (2015). An introduction to cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (eds), *An Introduction to Model-Based Cognitive Neuroscience*, pp. 3–24. New York: Springer.

Fink, B., Manning, J. T., & Neave, N. (2004). Second to fourth digit ratio and the 'big five' personality factors. *Personality and Individual Differences*, *37*, 495–503.

García-Pérez, M. A. (2005). On the confidence interval for the binomial parameter. *Quality & Quantity*, *39*, 467–481.

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, *3*:325.

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics*, *73*, 2332–2352.

García-Pérez, M. A., & Alcalá-Quintana, R. (2012). On the discrepant results in synchrony judgment and temporal-order judgment tasks: A quantitative model. *Psychonomic Bulletin & Review*, *19*, 820–846.

García-Pérez, M. A., & Alcalá-Quintana, R. (2015a). Converging evidence that common timing processes underlie temporal-order and simultaneity judgments: A model-based analysis. *Attention, Perception, & Psychophysics*, *77*, 1750–1766.

García-Pérez, M. A., & Alcalá-Quintana, R. (2015b). The left visual field attentional advantage: No evidence of different speeds of processing across visual hemifields. *Consciousness and Cognition*, *37*, 16–26.

García-Pérez, M. A., & Alcalá-Quintana, R. (2015c). Visual and auditory components in the perception of asynchronous audiovisual speech. *i-Perception*, 6(6), 1–20.

García-Pérez, M. A., & Núñez-Antón, V. (2013). Correlation between variables subject to an order restriction, with application to scientometric indices. *Journal of Informetrics*, *7*, 542–554.

García-Pérez, M. A., Núñez-Antón, V., & Alcalá-Quintana, R. (2015). Analysis of residuals in contingency tables: Another nail in the coffin of conditional approaches to significance testing. *Behavior Research Methods*, *47*, 147–161.

García-Pérez, M. A., & Peli, E. (2014). The bisection point across variants of the task. *Attention, Perception, & Psychophysics*, *76*, 1671–1697.

García-Pérez, M. A., & Peli, E. (2015). Aniseikonia tests: The role of viewing mode, response bias, and size–color illusions. *Translational Vision Science & Technology*, *4*(3):9, 1–22.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*, 2595–2602.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*, 647–674.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The *p* value fallacy. *Annals of Internal Medicine*, *130*, 995–1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, *130*, 1005–1013.

Goodman, S. N. (2005). Introduction to Bayesian methods I: Measuring the strength of evidence. *Clinical*

Trials, *2*, 282–290.

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143.

Hönekopp, J., & Watson, S. (2010). Meta-analysis of digit ratio 2D:4D shows greater sex difference in the right hand. *American Journal of Human Biology*, *22*, 619–630.

Hönekopp, J., & Watson, S. (2011). Meta-analysis of the relationship between digit ratio 2D:4D and aggression. *Personality and Individual Differences*, *51*, 381–386.

Hubbard, R. (2011). The widespread misinterpretation of *p*-values as error probabilities. *Journal of Applied Statistics*, *38*, 2617–2626.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*, 411–414.

Kaji, A. H., & Lewis, R. J. (2010). Are we looking for superiority, equivalence, or noninferiority? Asking the right question and answering it correctly. *Annals of Emergency Medicine*, *55*, 408–411.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*, *72*, 210–219.

Killeen, P. R. (2005a). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science*, *16*, 1009–1012.

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, *13*, 549–562.

Kline, R. B. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: American Psychological Association.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*, 16–26.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory and Psychology*, *22*, 67–90.

Lee, D. (2011). *Testing for a Poisson mixture: Comparison of the power of the posterior predictive check (PPC) and bootstrap approaches*. Doctoral Dissertation, Stony Brook University.

Leek, J. T., & Peng, R. D. (2015). *p* values are just the tip of the iceberg. *Nature*, *520*, 612.

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & Waldman, I. D. (2015). Fifty psychological and psychiatric terms to avoid: A list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Frontiers in Psychology*, *6*:1100.

Loh, W.-Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, *34*, 1818–1833.

Luce, R. D. (1985). Mathematical modeling of perceptual, learning, and cognitive processes. In S. Koch & D. E. Leary (eds), *A Century of Psychology as a Science*, pp. 654–677. New York: McGraw-Hill.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*, 1–26.

Manning, J. T., & Fink, B. (2008). Digit ratio (2D:4D), dominance, reproductive success, asymmetry, and sociosexuality in the BBC Internet Study. *American Journal of Human Biology*, *20*, 451–461.

Manning, J. T., Scutt, D., Wilson, J., & Lewis-Jones, D. I. (1998). The ratio of 2nd to 4th digit length: A predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Human Reproduction*, *13*, 3000–3004.

Meng, X.-L. (1994). Posterior predictive *p*-values. *Annals of Statistics*, *22*, 1142–1160.

Miller, G. A. (1965). Some preliminaries to psycholinguistics. *American Psychologist*, *20*, 15–20.

Millet, K., & Dewitte, S. (2006). Second to fourth digit ratio and cooperative behavior. *Biological*

Psychology, *71*, 111–115.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290.

Mueller, C. J., & Kuchinke, L. (2016). Processing of face identity in the affective flanker task: A diffusion model analysis. *Psychological Research*, in press.

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115..

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Nathoo, F. S., & Masson, M. E. J. (2016). Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology*, *72*, 144–157.

Ng, T.-H. (2014). *Noninferiority Testing in Clinical Trials: Issues and Challenges*. Boca Raton, FL: CRC Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Pardo, S. (2014). *Equivalence and Noninferiority Tests for Quality, Manufacturing and Test Engineers*. Boca Raton, FL: CRC Press.

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, *13*, 739–747.

Pernet, C. R., Wilcox, R., & Rousselet, G. A. (2013). Robust correlations analyses: False positive and power validation using a new open source Matlab toolbox. *Frontiers in Psychology*, *3*:606

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., & Evans, S. J. W. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *Journal of the American Medical Association*, *295*, 1152–1160.

Ranson, R., Stratton, G., & Taylor, S. R. (2015). Digit ratio (2D:4D) and physical fitness (Eurofit test battery) in school children. *Early Human Development*, *91*, 327–331.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Smith, P. (2015). Modeling simple decisions and applications using a diffusion model. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (eds), *The Oxford Handbook of Computational and Mathematical Psychology*, pp. 35–61. New York: Oxford University Press.

Ratcliff, R., & van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin & Review*, *16*, 742–751.

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain and Cognition*, *55*, 374–382.

Richards, G., Stewart-Williams, S., & Reed, P. (2015). Associations between digit ratio (2D:4D) and locus of control. *Personality and Individual Differences*, *83*, 102–105.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, 2002, *14*, 184–201.

Rothmann, M. D., Wiens, B. L., & Chan, I. S. F. (2011). *Design and Analysis of Non-Inferiority Trials*. Boca

Raton, FL: CRC Press.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283–300.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*, 411–432.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Smith, P. L., & Ratcliff, R. (2015). An introduction to the diffusion model of decision making. In B. U. Forstmann & E.-J. Wagenmakers (eds), *An Introduction to Model-Based Cognitive Neuroscience*, pp. 49–70. New York: Springer.

Sridharan, D., Steinmetz, N. A., Moore, T., & Knudsen, E. I. (2014). Distinguishing bias from sensitivity effects in multialternative detection tasks. *Journal of Vision*, *14*(9):16, 1–32.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

Trafimow, D., & Marks, M. (2016). Editorial. *Basic and Applied Social Psychology*, *38*, 1–2.

Tunes da Silva, G., Logan, B. R., & Klein, J. P. (2009). Methods for equivalence and noninferiority testing. *Biology of Blood and Marrow Transplantation*, *15*, 120–127.

Turner, B. M., Sederberg, P. B., & McClelland, J. L. (2016). Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*, *72*, 191–199.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after NHST: How to describe your data without "*p*-ing" everywhere. *Basic and Applied Social Psychology*, *37*, 260–273.

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General*, *143*, 1794–1805.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (eds.), *The Oxford Handbook of Computational and Mathematical Psychology*, pp. 300–319. New York: Oxford University Press.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*, 385–402.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (in press). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. Waldman (eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*. New York: Wiley.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd edition). Boca Raton, FL: CRC Press.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*, 1293–1313.

Yang, H., Meijer, H. G. E., Doll, R. J., Buitenweg, J. R., & van Gils, S. A. (2015). Computational modeling of Adelta-fiber-mediated nociceptive detection of electrocutaneous stimulation. *Biological Cybernetics*, *109*, 479–491.