Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

NULL HYPOTHESIS SIGNIFICANCE TESTING IN SIMULATION

Marko A. Hofmann

ITIS
University of the Federal Armed Forces Munich
Werner-Heisenberg-Weg 39, Neubiberg 85577, GERMANY

ABSTRACT

Several papers have recently criticized the use of null hypothesis significance testing (NHST) in scientific applications of stochastic computer simulation. Their criticism can be underpinned by numerous articles from statistical methodologists. They have argued that focusing on *p*-values is not conducive to science, and that NHST is often dangerously misunderstood . A critical reflection of the arguments contra NHST shows, however, that although NHST is indeed ill-suited for many simulation applications and objectives it is by no means superfluous, neither in general, nor in particular for simulation.

1 INTRODUCTION AND MOTIVATION

In the behavioral and social sciences as well as in medicine null hypothesis significance testing (NHST) has dominated statistical reasoning for decades. Even in the natural and technical sciences most statistical university lecturers spent hours of teaching on this method. However, the use of significance testing in the analysis of research data has been criticized, both logically and conceptually, from numerous statisticians – continuously for almost 100 years: Boring 1919; Berkson 1938; Bakan 1966; Greenwald 1975; Carver 1978; Rosnow and Rosenthal 1989; Tukey 1991; Cohen 1994; Sedlmeier 1996; Schmidt and Hunter 1997; Falk 1998; Gigerenzer 2004; Ioannidis 2005; Hubbard and Armstrong 2006; Armstrong 2007; Hubbard and Lindsay 2008; Lambdin 2012; Lin, Lucas, and Shmueli 2013; and Schneider 2015. Some of the suggestions for improvement have been summarized by Cumming (2014) who coined the term "New Statistics" for a general and categorical shift from NHST mainly towards confidence intervals and statistical effect sizes. Now, the criticism has been given an almost official status by a statement issued on March 7, 2016 from the American Statistical Association on statistical significance and *p*-values (Wasserstein and Lazar 2016). Recently, White et al. (2014) and Troitzsch (2014) have criticized the use of NHST in the context of the simulation-based evaluation of research hypotheses *"scientific applications of computer simulation*, too. Both articles recommend to focus on *effect sizes* instead of *p*-values for the interpretation of simulation results. In (Hofmann 2015b) I have advocated for the replacement of *p*-values by confidence intervals of effect sizes whenever simulation results are, for a first overview, condensed into single values. The article at hand tries to summarize the current state of criticism on NHST in general and with respect to computer simulation in a more balanced way.

1.1 Deficits of significance tests and counterarguments

A perfectly executed summary of the numerous pros and cons of NHST has been given by Nickerson (2000) (in an article of 60 pages!). Up-to-date critical summaries can be found in (Lambdin 2012, Kline 2013, Cumming 2014, Schneider 2015). A recent defense has been published by Morey et al. (2014). Important fundamental defenses can be found in (Dooling and Danks 1975; Mulaik, Raju, and Harshman 1997; Cortina and Dunlap 1997; Hagen 1997; Senn 2001) and, implicit, in (Miller 2009). (Most of these articles have been published in psychological journals. The reason is that inferential statistics is seen

as the essential backbone of scientific reasoning in the behavioral sciences, at least for the last 60 years (Gigerenzer et al. 1989).

It is beyond the limits of this article to discuss all the arguments discussed in these papers. Yet, a closer look at the the major deficits and counterarguments might sensitize the simulation practitioner that inferential statistics is far from being the non-controversial issue taught in standard textbooks. For a short introduction into NHST see (Hofmann 2015b). The counterarguments should not be misunderstood as final refutations of the deficits. It would be easy for the advocates and opponents of NHST to start long disputations about all the topis mentioned.

1. *Deficit:* The central assumption of NHST, that there is absolutely no difference between the means of two groups ($\mu_1 = \mu_2$) is basically a straw man. In almost all research questions groups will have different means if we measure precisely ("They are always different - for some decimal place" Tukey (1991), p. 100)). The crucial question is whether this difference is practically important or not. Law (2014) used the same argument in order to explain why NHST is inadequate for simulation model validation: "Since the model is only an approximation to the actual system, a null hypothesis that the system and model are the 'same' is clearly false (p. 269)".
   Counterarguments: (a) The criticism is based on *nil* hypotheses ("no effect") which are a special sub-class of *null* hypotheses ("hypotheses to be nullified"). A null hypothesis, in general, can be any point estimate or interval. Thus, one can tailor the statistical null hypothesis to any research hypothesis one has, including a tolerated deviation from the "no-effect-hypothesis". If a null hypothesis is true, however, NHST is not affected by sample size: In that case the p-value cannot be reduced to significant levels only by increasing sample size. Tests that are able to cope with general research hypothesis are: Non-nil null hypothesis, in general (Nickerson 2000, p. 274), equivalence tests (Parkhurst 2001), minimum-effect-tests (Murphy and Myors 1999), and, especially for simulation applications, interval hypothesis tests (Sargent, Goldsman, and Yaacoub 2015). All these tests stay within the general logic of NHST; they directly render irrelevant the technically correct reasoning of Law (2014) against NHST for model validation, and, indirectly, refute this standard reasoning against NHST itself: Since any research hypothesis (point or interval, including specific minimal effects) can be formulated as a statistical null hypothesis, and since p-values are not affected by sample size if the null hypothesis is true, a p-value can make perfect sense even for huge (simulated) samples. (b) Frick (1996) pointed out that there are important cases of research for which it is reasonable to treat the point null hypothesis as true. (c) (Krueger 2001, p.19) stated that the association between variables "might well be ridiculously small, but a judgment about the ridiculousness of an effect size is not part of NHST. This judgment can be made only by a human appraising the size of the effect and the size of the sample necessary to coax this effect into significance." (d) A null hypothesis can be the most extreme point hypothesis from the set of all hypothesis one would like to reject. This reasoning is obvious in an one-sided test: If the researcher assumes, for example, that $\mu > 0$, then the most conservative test is to assume $\mu = 0$. (Since one-sided tests can be easily misused (Lombardi and Hurlbert 2009b) it is recommended to apply them with great care.)
2. *Deficit:* NHST applies the *modus tollens* ("If P then Q; not Q; therefore not P") on probabilistic statements (If H_0 is true then probably $p > 0.05$; $p < 0.05$; therefore probably H_0 is false). Berkson (1942) was the first author who criticized this application as faulty. Cohen (1994) demonstrated that the transfer from categorical premises to probabilistic ones is logically wrong using an example taken from (Pollard and Richardson 1987): "If a person is an American, then he is probably not a member of Congress. (TRUE, RIGHT?) This person is a member of Congress. Therefore, he is probably not an American (Cohen 1994, p. 998)." This proposition is formally exactly the same as "If H_0 is true, then this result (statistical significance) would probably not occur. This result has occurred. Then H_0 is probably not true and therefore formally invalid."

Counterargument: The example used by Cohen is rather extreme. Cortina and Dunlap (1997) showed that in most cases the modus tollens can be successfully applied even in a probabilistic context. Their reasoning is similar to the fundamental defense Krueger (2001) has given: Inductive inference cannot be justified on the basis of binary logic, but it can be defended pragmatically. With other words: Although logically flawed, "much has been learned from NHST in the past and presumably more can be learned in the future (Krueger 2001, p. 23)."

3. *Deficit:* NHST is based on indirect reasoning: The likelihood of a research hypothesis (H_1) is indirectly assessed via the conditional probability $Pr(D|H_0)$ to get the observed sample data D, assuming the contrary hypothesis H_0. p-values are therefore extremely easy to misinterpret for non-experts (Goodman 2008, Lambdin 2012) as well as experts (Haller and Krauss 2002). A p-value, to specify three of the most popular errors, is neither the probability that the results obtained occurred due to chance nor the probability of the null hypothesis given the data (= $Pr(H_0|D)$) nor does a p-value of 0.05 imply that we have observed data that would occur only 5% of the time under the null hypothesis. It is the probability $Pr(D+|H_0)$ to get the observed data (D) *or more extreme results* ($D+$), assuming the null hypothesis H_0 is true in the population (even more precisely: $p := Pr(T(X) > T(D)|H_0)$; where $T(X)$ is the specific test statistic (Lambdin 2012)).
   Counterarguments: (a) Learning the formally correct interpretation of a p-value is a matter of proper teaching. (b) Much of the confusion is generated by ambiguity and imprecision of casual language (Nickerson 2000, p. 262) which is ill-suited for mathematical concepts like conditional probabilities. One should simply avoid to much prose. Just report the exact p-value (Greenwald et al. 1996).
4. *Deficit:* The probability $Pr(D+|H_0)$ expressed as a p-value does not tell the researchers what they want to know. They are interested in the probability $Pr(H_0|D)$ or even better in a measure that indicates whether and how much H_1 is more likely than H_0. Such results are only calculable with Bayesian inference using a priori probabilities (Kline 2013, Lecoutre and Poitevineau 2014).
   Counterarguments: (a) Fisher's evidential statistics is based on a sensible argument by contradiction: If the contrary to the research hypothesis is assumed to be true and the data are very unlikely under this assumption, then we undeniably have some evidence for the research hypothesis. In the context of scientific inquiry (in contrast to practical decision making) it is essential to treat the p-values of single studies not as hard numbers but as indicators (Greenwald et al. 1996). (b) Nickerson (2000), p. 252) showed that if one can assume that $Pr(H_1)$ is at least as large as $Pr(H_0)$ and that $Pr(D|H_1)$ is much larger than $Pr(D|H_0)$ (assumptions that should apply in many if not most cases), then a small value of $p = Pr(D|H_0)$ can be take as a proxy for a relatively small value of $p = Pr(H_0|D)$. (c) On the one hand, Bayesian inferential statistics can avoid many of the disadvantages of NHST (Dienes 2011; Dienes 2014; Rouder 2014; Wagenmakers et al. 2015). A direct mathematical comparison of "Fisher, Neyman-Person, and Bayes" from Christensen (2005) clearly favored Bayes. On the other hand, a change to Bayesian inference necessitates an extensive reform of statistical education without a guarantee that it will fix the problem of incorrect applications and conclusions from laymen, and it has its critics, too: Lombardi and Hurlbert (2009b), p. 465) claim that "the literature on Bayesian statistics has a surfeit of inclarities and errors ... Problems include: the usual widespread confusion over the distinction between statistical hypotheses and scientific hypotheses; failure to recognize that inferential methods in most disciplines are useful mainly, albeit implicitly, for purposes of estimation; a focus on uncommon types of testing situations, such as a comparison of a point null and a point alternative hypothesis; claims of 'objectivity' based on highly biased prior probabilities, as when the H_0 of 'no effect' is assigned a prior of 0.5; and, most generally, the injection of more subjectivity into statistical analyses before interpretation of results of such analyses." Moreover, the calculation of Bayesian factors from samples is theoretically only as powerful as the calculation of exact p-values. The probability of replicating a statistical effect from a given sample, for example, is not affected by the type of inference applied. In that sense,

the different ways of summarizing a given set of sample data are informationally equivalent (Miller 2009). Savalei and Dunn (2015) therefore concluded: "...at present we lack empirical evidence that encouraging researchers to abandon p-values will fundamentally change the credibility and replicability of psychological research in practice." Finally, the recommendation to replace NHST by Bayesian inference has already been made 50 years ago in a famous article by Bakan (1966) - without any recognizable effect (In a recent study (Yu et al. 2014) only 1 of the 314 scientists who responded to a questionnaire about statistically supported decision making reported using Bayesian methods). Therefore, although Bayesian inference benefits specifically from new statistical software solutions (Kruschke 2015) it is not to be expected that Bayesian thinking will triumph very soon. Due to this academic inertia even a fervent advocate of Bayesian inference like Goodman (2008) is skeptical about a paradigmatic shift, and recommends to consider confidence intervals and effects sizes "in the meantime". The same pessimism has been stated by Christensen (2005).

5. *Deficit:* The probability $Pr(D+|H_0)$ includes not only the observed data but also unobserved more extreme data. R. A. Fisher (the "godfather" of significance testing; (Fisher 1925, Fisher 1935)) said he threw in the rest of the tail area "as an approximation (Goodman 2008, p. 137)". There is no logical justification for such an inclusion. Bayesian inference, for example, works without it.
   Counterargument: An athlete has jumped 9.01 meters. One would not put into question his performance by asking: "How likely is it that a doping-free athlete in the long jump manages to jump exactly 9.01 meters?" One would ask: "How likely is it that a doping free athlete exceeds 9.00 meters?"
6. *Deficit:* Since even the smallest effect is significant above a certain sample size, it is only a matter of effort to demonstrate significance (Berkson 1938). Consequently, a foolproof guideline for demonstrating significance is to apply a flexible stopping rule for sample size n: Increase it until p-value < 0.05 then stop and publish (Simmons, Nelson, and Simonsohn 2011). This is the reason why NHST is indeed completely inadequate for *exploratory* research based on simulation: Since increasing sample size with closed (non-interactive) simulations is trivial, NHST is next to meaningless for the comparison of purely simulated data sets (with no direct backup from empirical samples). The recent criticism of White et al. (2014) and Troitzsch (2014) on NHST as the standard statistical evaluation tool for computer simulations is based on this deficit.
   Counterarguments: (a) NHST is not intended for the exploratory research mentioned in (White et al. 2014) and (Troitzsch 2014) at all. (Scientific investigation are called exploratory if they cannot be sufficiently grounded on empirical data; they are in contrast to evaluative or confirmatory investigations, based on empirically validated models). The application of any kind of inferential statistics on data generated by exploratory simulation is nonsensical. Empirical and exploratory-simulation-generated data sets cannot be equalized. The adequate method of statistical reasoning for exploratory simulation results is, as the name suggests, exploratory data analysis (EDA) (Tukey 1977; Hoaglin, Mosteller, and Tukey 2000). EDA renounces inference in favor of descriptive methods like box-plot or stem-and-leaf diagrams. The main advantage of EDA is its proximity to the raw data (approaching the ideal of the complete set of statistical data), and the information density provided by a *good* graphic (Tufte 2001, Few 2009). An excellent guideline for presenting quantitative data in general has been given by Gillan et al. (1998). (b) Pseudo random number generators (PRNGs) are ideal examples for the sucessful application of NHST in simulation. Basically, PRNGs are approximate simulations of true random processes (TRP; atmospheric noise or radioactivity, for example). Huge samples (up to 10^8!) are considered to be most adequate for testing PRNGs (Soto 1999, Haramoto 2009, Rukhin et al. 2010, L'Ecuyer 2015) – as long as the samples are not larger than the total number of different numbers generated by the PRNG (its period). Only extremely large samples have enough power to find the tiny effects that can reveal a bias in modern PRNGs. With respect to such specific attributes like independence, goodness-of-fit, or lengths of runs a PRNG is only acceptable if it is not possible to demonstrate a significant deviation from a

true random process even with huge samples. Obviously, PRNGs are practical counterexamples against the standard reasoning against NHST in simulation: Although PRNGs are not equal to TRP for theoretical reasons (the nil hypothesis of randomness is definitely false) they are tested via significance tests based on huge samples generated by algorithms that mimic (simulate) randomness. The trick is simply to consider p-values down to 10^{-4} as inconclusive. A PRNG is finally rejected only if the p-values are as low as 10^{-8} (Haramoto 2009). (c) You cannot blame the high statistical power $(1-\beta)$ caused by huge sample size of being able to detect even tiny effects. This is exactly the reason why we want to have high power! (d) Flexible stopping rules are questionable research practice in general, they are not specific to NHST.

7. *Deficit:* Significance tests are almost useless when sample size is small, since only huge effects in small samples (visible at first sight) generate p-values which cause the null hypothesis to be rejected. All other effects lead to non-rejection, which does neither provide evidence for nor against the null hypothesis. Although it is true that in small samples the probability is high that an effect is caused by sample error, the p-value alone can entirely hide even a substantial effect in a small sample. From a logical point of view, one should perform equivalence tests ($H_0 \sim |\mu_2 - \mu_1| \geq \Delta$) in such situations (Parkhurst 2001). The null hypothesis in such tests in not a point hypothesis but that the difference is not negligible (exceeds a predefined threshold Δ). Unfortunately, only experts seem to know about them, although equivalence tests are conceptually of equal importance to science (including simulation-based science) than significance testing and often easy to perform without further effort (Neuhaeuser 2010). Equivalence tests have even been proposed as the ideal statistical method of model *validation* (Robinson and Froese 2004).
   Counterargument: Equivalence tests are ultimately based on exactly the same logic as significance tests. They are an useful extension to NHST, not a challenge (see deficit 1). Simplified version of such test that stay completely within the methodology of NHST have been proposed by Serlin and Lapsely (1985) and Murphy and Myors (1999). The latter claim that "the biggest single obstacle to adopting the approach described her is that consensus must be reached about a working definition for negligible effects", and not any reform of NHST.
8. *Deficit:* Due to the rather arbitrary 5% rule of rejection (p-value < 0.05), and the tendency of journals to publish significant results only, NHST have caused a serious "file drawer" problem (Rosenthal 1979, Fanelli 2012): Non significant studies do not get published, although each of them is an important contribution to science. This disregard is more critical than it might seem at first sight, since a series of non-published, non-significant studies could even uncover a strong effect in meta-analysis (Thompson 2007, p. 429). The problem is also discernible in computer simulation: I have witnessed several data farming experiments (Hofmann 2013) in which all non-significant results were immediately discarded. The deeper reason for this imbalance seems to be a strong bias against the null hypothesis (Greenwald 1975) inherent to the method of NHST: Since non-rejection is considered to be inconclusive only the rejection of the null hypothesis is appreciated. "The consequence is an ego involvement with rejection of the null hypothesis that often leads researchers to interpret null hypothesis rejections as valid confirmations of their theoretical beliefs while interpretating non-rejection as uninformative and possibly the result of flawed methods.(Greenwald et al. 1996, p. 177)."
   Counterarguments: (a) The general 5% rule is a (bad) human choice in some research fields. It is not a inherent problem of NHST. (b) Non-significant studies do not get published because journal space is rare (Nosek, Spies, and Motyl 2012). A change towards less selective online portals instead of printed journals will solve the problem (Eve 2012).
9. *Deficit:* Stand alone, significance test put far too much emphasis on the probability of a random effect, without telling anything about the magnitude of the effect in the sample. The magnitude of an effect is, however, at least in general, the very first thing one would like to know about (Hofmann

2015b). For the purpose of assessing the practical significance, effect sizes are the measures of first choice not p-values (Ellis 2010, Grissom and Kim 2012, Kelley and Preacher 2012, Kline 2013).
Counterarguments: (a) Effect sizes can be added to p-values (Wilkinson 1999). (b) The first question should be: Is there a true effect or an artifact created by randomness? If this question is answered in favor of the effect then the size of it becomes interesting.

10. *Deficit:* Similarly, p-values put far too much emphasis on an exact single value. They do not communicate the uncertainty introduced by random sampling. They share this drawback with effect sizes which are only point estimates. The standard statistical measure of precision/uncertainty for estimates are confidence intervals. Their range is an indispensable additional information (Cumming 2014).
Counterarguments: (a) Confidence intervals can be added to p-values (Wilkinson 1999). (b) CI are as easily misinterpreted as p-values (Belia et al. 2005; Hoekstra, Morey, and J. N. Rouder 2014). (c) The width of a CI also depends on sample size.
11. *Deficit:* The difference between a "significant" and a "non-significant" result is not itself significant (Gelman and Stern 2006), regardless of the specific level of significance ($p < .05$ or $p < .01$). One cannot hardly claim that there is an important difference if the p-values from treatment groups A and B (calculated with respect to a control group K) are .049 and .051, for example. The error, however, seems quite common: In a study on five top-ranking journals including Nature and Science 79 from 157 articles that made such comparisons applied the wrong reasoning (Nieuwenhuis, Forstmann, and Wagenmakers 2011).
Counterargument: In making a comparison between two different groups, treatments, or parameter settings (for simulation), one has to look at the statistical significance of the difference rather than the difference between their significance levels. The wrong reasoning is not a deficit of NHST. It is simply a misapplication.
12. *Deficit:* The interpretation of several independent study results on the sole basis of p-values is illogical: Suppose that the first study (for details see (Cumming 2011, p. 1-14)) summarizes the results by the p-value $= .02$ whereas the second study is concentrated into the p-value $= .22$. Thus, one study result is significant, the other clearly not. What is the standard conclusion within the logic of NHST? According to Cumming (2011) many researchers find such a situation "inconsistent" or "equivocal". Further research is required for any interpretation. However, the most justifiable interpretation is "consistent", since both studies demonstrate an effect into the same direction. If NHST is further used to test whether there is any difference between the two studies one gets the p-value .55 which is really as close to perfect similarity as could be reasonably be expected from two independent studies. The p-value of the combined studies would be .008!
Counterargument: Again, the wrong reasoning is, obviously, not a deficit of NHST. It is a misinterpretation based on the dichotomous thinking of a predefined α-level of rejection. It is a perfect example for the reasons why R. A. Fisher (p-values) did not like the Neyman-Pearson approach (α and β errors) to significance testing.
13. *Deficit:* Most of the standard tests of NHST (t-test, F-test etc.) are non-robust towards outliers and violation of normality, equal variance and equal size in the samples (Wilcox 2012). Normality, however, seems to be a rather rare distribution in reality (Micceri 1989), as well as outliers seem to be rather frequent, a result predicted in a seminal paper by Tukey (1960).
Counterargument: Robust tests (Wilcox 2012) have been developed within the paradigm of NHST. They are based primarily on trimming, winsorizing and bootstrapping samples. In addition, one would like to recommend randomization tests (Edgington and Onghena 2007), which are basically (almost) assumption-free, exact tests based on complete or partial permutations of the original sample data. Both improvements are relatively easy to include into the standard statistical education: Lock et al. (2013), for example, have written an introductory textbook for inferential statistics based on randomization tests. Most simulation practitioners will benefit from these techniques without

much difficulties. For confirmatory research (as opposed to exploratory research) based on precise physical or technical simulation both test variants have many advantages.

14. *Deficit:* p-values say nothing about the replicability of a study result (Lykken 1968, Falk and Greenbaum 1995).
    Counterarguments: (a) Greenwald et al. (1996), 179 showed that (under reasonable assumptions) " a p-value resulting from NHST is monotonically related to an estimate of a non-null finding's replicability." (b) Killeen (2005) used a transformation of p-value to calculate p_{rep}, a probability measure of replicability. (c) Miller (2009) proved that it is unfair to regard this deficit as indicative of a problem unique to NHST. The estimations for replication probabilities are of little diagnostic value, *regardless of the method applied to calculate them*.
15. *Deficit:* NHST has misled countless researches into rigid mechanical reasoning based on the simplistic p-value $< \alpha$ decision rule (Rosnow and Rosenthal 1989, Lambdin 2012) and its dichotomous thinking. Although refutation or corroboration are central goals of science, they are not apodictic verdicts based on binary values. Magnitude, precision and likelihood of effects have to be assessed in context of domain, newness and impact.
    Counterarguments: (a) Dichotomous answers ideally satisfy the need to decide and act (Greenwald et al. 1996). One might only question whether science and decision making have the same goals (Hofmann 2015a). Anyway, for simulation-based methods like data farming which generate millions of samples some automated reasoning (based on thresholds) is indispensable. (b) Effect sizes and confidence intervals can (and should) be added to p-values (Wilkinson 1999).
16. *Deficit:* The famous $p < \alpha$-rule combines Fisher's evidential statistic (p-value) and Neyman-Pearson's error estimate approach (α and β) in a way not justified by either theory (Hubbard 2004, Gigerenzer 2004, Schneider 2015).
    Counterarguments: (a) Depending on the context it is possible to use either Fisher's evidential statistics or Neyman-Pearson's error estimate approach exclusively. (b) The two theories are (mathematically) closer to each other as it is often claimed (Lehmann 1993). (c) The "null ritual" (Gigerenzer 2004) based on the $p < \alpha$-rule is simply a misuse of both Fisher's evidential statistic and Neyman-Pearson's error estimate approach. It is a wrong way of doing NHST, not a fault of NHST.
17. *Deficit:* NHST seems ill-suited to counter the human tendency to sacrifice soundness of argumentation for the opportunity to publish. This has lead to alarming analyzes about erroneous scientific reports. The best-known work in that direction is "Why most published research findings are false" from Ioannidis (2005). Although Ioannidis proved that high error rates in science are, at least up to a certain degree, unavoidable (due to limits in achievable sample size), the problem is aggravated by the methodology of NHST which allows "undisclosed flexibility" with respect to producing significant results (Simmons, Nelson, and Simonsohn 2011). To put it into simple but plain words: In some fields of science p-values have become the main force of selection (via high rejection rates of top journals) in a Darwinian process of academic success. Humans proved themselves to be highly adaptive to this force while ignoring truth and sincerity.
    Counterarguments: (a) There is no mathematical solution to counter the problems of "undisclosed flexibility" and "publication bias" in the Darwinian process (McQuarrie 2014) of "publish or perish". It is an socio-cultural issue that affects all statistical methods. Simmons, Nelson, and Simonsohn (2011) see the only solution in working discipline from authors and reviewers. They explicitly call four technical suggestions (correcting alpha levels, Bayesian statistics, replications, and posting data) "non-solutions". McQuarrie (2014) recommends to abolish journal scarcity. Nosek, Spies, and Motyl (2012) see the "ultimate solution" in categorical openness: They demand total transparency of scientific data, methods, tools, and workflow as well as the complete removal of barriers for publication. (b) p-values are not the decisive criteria to get published in computer simulation. The problem only exists in the human and social sciences.

As a final remark, note again that the counterarguments should not be considered as final rejections. Defenses and attacks of the arguments have been given and extended in numerous articles: Nickerson (2000) listed 400 papers on the topic, the majority of them critical about NHST; Schneider (2015) added about 50 newer (2000-2014) references. In the numerous month of reading a substantial part of these papers I learned that the more you know about the topic the less secure you become about the proper conclusion. There is only one point on which most writers agree: "NHST cannot be done mechanically without running the risk of obtaining nonsensical results (Nickerson 2000, p. 290)." I would like to add that NHST has proven itself hard to be refuted on logical or mathematical grounds using an uncontradicted line of reasoning.

2 CONCLUSIONS FOR SIMULATION

With special respect to simulation two strong conclusions can be drawn:

1. *p*-values are ill-suited for all exploratory simulation for three reasons: First (see deficit 6 and counterargument (a)), a researcher can achieve statistical significance by increasing effect size or sample size (Test statistic significance $\approx$ size of effect $\times$ sample size; (Maxwell and Delaney 1990)). Since increasing sample size in an exploratory setting is trivial only the effect size can be of any interest. Second, exact *p*-values create an illusion of precision which is seldom, if at all, justified in exploratory simulation. Third, and most important, equalizing samples from exploratory simulations models with empirical samples is self-deceiving. Exploratory simulation does not produce facts. Yet, all methods of inferential statistics are therefore ill-suited for exploratory research. Exploratory data analysis (EDA) (Tukey 1977; Hoaglin, Mosteller, and Tukey 2000) is much better for exploration – as the name also suggest. One will find the same skepticism towards inferential statistics and praise of EDA in the first publications on data farming (Horne and Meyer 2005). The authors emphasize that the method has the ability "to discover trends and outlier in results [1082]" and "to process large parameter spaces, makes possible the discovery of surprises (both positive and negative) and potential options [1082]". Unfortunately, current usage often seems to neglect the restrictions of validity in exploratory computer simulation (Hofmann 2013).
2. Nothing serious impedes the application of NHST if the simulation-generated data can be equalized with empirical data. I see three fields of application that fulfill this precondition: physical or technical simulations with great predictive validity, mathematical applications like PRNGs, and all internal testing of simulation systems before the link into a reference system is made. *p*-values can be an important measurement (among others) for a scrutiny in a confirmatory setting if the central question you have is whether the parameter-variation in the simulation model has caused an output change which is hardly explainable by sampling error alone. For *this* purpose Fisher's framework of evidential statistics seems more appropriate than Neyman-Pearson's error estimate approach, which is better suited for quality control in industrial production or comparable simulation applications.

A last questions remains unanswered: Are there any invulnerable alternatives to *p*-values? The answer depends on context and personal preferences. In some contexts effect sizes (ES) and confidence intervals (CI) are more informative than *p*-values (Hofmann 2015b), and Bayesian factors are much more informative *if* you have good priori probabilities. Yet, neither ES and CI nor Bayesian factors are indisputable (Hoekstra, Morey, and J. N. Rouder 2014; Lombardi and Hurlbert 2009a). Even in the recent criticism of *p*-values from the ASA (Wasserstein and Lazar 2016) you will (therefore) not find a general substitute for *p*-values.

REFERENCES

Armstrong, J. S. 2007. "Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries". *Internation Journal of Forecasting* 23:335–336.

Bakan, D. 1966. "The test of significance in psychological research". *Psychological Bulletin* 66:423–437.

Belia, S., F. Fidler, J. Williams, and G. Cumming. 2005. "Researchers misunderstand confidence intervals and standard error bars". *Psychological Methods* 10 (4): 389–396.

Berkson, J. 1938. "Some difficulties of interpretation encountered in the application of the Chisquare test". *Journal of the American Statistical Association* 33:526–536.

Berkson, J. 1942. "Tests of significance considered as evidence". *Journal of the American Statistical Association* 37:325–335.

Boring, E. 1919. "Mathematical vs. scientific significance". *Psychological Bulletin* 16:335–338.

Carver, R. 1978. "The case against stastistical significance testing". *Harvard Educational Review* 48:378–399.

Christensen, R. 2005. "Testing Fisher, Neyman, Pearson, and Bayes". *The American Statistician* 59 (2): 121126.

Cohen, J. 1994. "The earth is round ($p < 0.5$)". *American Psychologist* 12:997–1003.

Cortina, J. M., and W. P. Dunlap. 1997. "On the Logic and Purpose of Significance Testing". *Psychological Methods* 2 (2): 161–172.

Cumming, G. 2011. *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. London: Routledge.

Cumming, G. 2014. "The new statistics: why and how". *Psychological Science* 25:7–29.

Dienes, Z. 2011. "Bayesian versus orthodox statistics: Which side are you on?". *Perspectives on Psychological Science* 6 (3): 274–290.

Dienes, Z. 2014. "Using Bayes to get the most out of non-significatnt results". *Frontiers in Psychology* 5:1–17.

Dooling, D. J., and J. H. Danks. 1975. "Going Beyond Tests of Significance: Is Psychology Ready?". *Bulletin of the Psychonomic Society* 5 (1): 15–17.

Edgington, E. S., and P. Onghena. 2007. *Randomization Tests, 4th ed.* Chapman and Hall.

Ellis, P. D. 2010. *The essential guide to effect sizes*. Cambridge, GB: Cambridge University Press.

Eve, M. P. 2012. "Tear it down, build it up: the research output team, or the library-as-publisher". *Insights UKSG* 25:158162.

Falk, R. 1998. "In criticism of the null hypothesis statistical test". *American Psychologist* 53:798–799.

Falk, R., and C. W. Greenbaum. 1995. "Significance tests die hard". *Theory & Psychology* 5:75–98.

Fanelli, D. 2012. "Negative results are disappearing from most disciplines and countries". *Scientometrics* 90 (3): 891–904,.

Few, S. 2009. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press.

Fisher, R. A. 1925. *Statistical methods for reseach workers*. London: Oliver & Boyd.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Frick, R. W. 1996. "The appropriate use of null hypothesis testing". *Psychological Methods* 1:379–390.

Gelman, A., and H. Stern. 2006. "The difference between 'significant' and 'not significant' is not itself statistically significant". *The American Statistician* 60:328–331.

Gigerenzer, G. 2004. "Mindless statistics". *The Jounal of Socio-Economics* 33:587–606.

Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Krueger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. New York: Cambridge University Press.

Gillan, D. J., C. D. Wickens, J. G. Hollands, and C. M. Carswell. 1998. "Guidellines for presenting qualitative data in HFES publications". *Human Factors* 40:28–41.

Goodman, S. N. 2008. "A dirty dozen: Twelve P-value misconceptions". *Seminars in Hematology* 45 (3): 135–140.

Greenwald, A. 1975. "Consequences of predjudice agains the null hypothesis". *Psychological Bulletin* 82:1–20.

Greenwald, A. G., R. Gonzales, R. J. Harris, and D. Guthrie. 1996. "Effect sizes and p values: What should be reported and what should be replicated?". *Psychophysiology* 33:175–183.

Grissom, R., and J. Kim. 2012. *Effect Sizes of Research*. Routledge.

Hagen, R. 1997. "In praise of the null hypothesis test". *American Psychologist* 52:15–24.

Haller, H., and S. Krauss. 2002. "Misinterpretations of significance: A problem students share with their teachers?". *Methods of Psychological Research Online* 7 (1): 1–20.

Haramoto, H. 2009. "Automation of Statistical Tests on Randomness to Obtain Clearer Conclusion". In *Monte Carlo and Quasi-Monte Carlo Methods 2008*, edited by P. L' Ecuyer and A. B. Owen, 411–421. Springer Berlin Heidelberg.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 2000. *Understanding Robust and Exploratory Data Analysis*. Wiley.

Hoekstra, R., R. Morey, and E. J. W. J. N. Rouder. 2014. "Robust misinterpretation of confidence intervals". *Psychonomic Bulletin & Review* 21 (5): 1157–1164.

Hofmann, M. 2013. "Simulation-based exploratory data generation and analysis (data farming): a critical reflection on its validity and methodology". *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 10 (4): 381–393.

Hofmann, M. 2015a. "Reasoning beyond predictive validity: The role of plausibility in decision-supporting social simulation". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti. Piscataway, New Jersey: IEEE.

Hofmann, M. 2015b. "Searching for effects in big data: Why p-values are not advised and what to use instead". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti. Piscataway, New Jersey: IEEE.

Horne, G. E., and T. E. Meyer. 2005. "Data Farming: Discovering Surprise". In *Proceedings of the 2005 Winter Simulation Conference*, edited by F. B. A. M. E. Kuhl, N. M. Steiger and J. A. Joines. Piscataway, New Jersey: IEEE.

Hubbard, R. 2004. "Alphabet soup: Blurring the distinctions between ps and alphas in psychological research". *Theory & Psychology* 14:295327.

Hubbard, R., and J. Armstrong. 2006. "Why we don't really know what statistical significance means: Implications for educators.". *Journal of Marketing Education* 28:114–120.

Hubbard, R., and R. M. Lindsay. 2008. "Why p values are not a useful measure of evidence in stastistical significance testing". *Theory and Psychology* 18:69–88.

Ioannidis, J. 2005. "Why most puplished research findings are false". *PLoS Medicine* 2 (8): e124.

Kelley, K., and K. J. Preacher. 2012. "On effect size". *Psychological Methods* 17 (2): 137–152.

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

Kline, R. B. 2013. *Beyond Significance Testing : Statistics Reform in the Behavioral Sciences (2nd edition)*. American Psychological Association.

Krueger, J. 2001. "Null hypotheses significance testing: On the survival of a flawed method". *American Psychologist* 56 (1): 16–26.

Kruschke, J. K. 2015. *Doing Bayesian Data Analysis (2nd ed.)*. Academic Press.

Lambdin, C. 2012. "Significance tests as sorcery: Science is empirical - significance tests are not". *Theory and Psychology* 22 (1): 67–90.

Law, A. M. 2014. *Simulation Modeling and Analysis (5th ed.)*. Mcgraw-Hill.

Lecoutre, B., and J. Poitevineau. 2014. *The Significance Test Controversy Revisited*. Springer.

L'Ecuyer, P. 2015. "Random number generators with multiple streams for sequential and parallel computing". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti. Piscataway, New Jersey: IEEE.

Lehmann, E. L. 1993. "The Fisher, Neyman-Pearson theories of testing hypotheses: One theoriy or two?". *Journal of the American Statistical Association* 88 (424): 1242–1249.

Lin, M., H. Lucas, and G. Shmueli. 2013. "Research Commentary -Too Big to Fail: Large Samples and the p-Value Problem". *Information Systems Research* 24 (4): 906–917.

Lock, R., P. F. Lock, K. L. Lock, E. F. Lock, and D. F. Lock. 2013. *Statistics: Unlocking the Power of Data*. Hoboken, NY: Wiley.

Lombardi, C. M., and S. H. Hurlbert. 2009a. "Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian". *Annales Zoologici Fennici* 46:311–349.

Lombardi, C. M., and S. H. Hurlbert. 2009b. "Misprescription and misuse of one-tailed tests". *Austral Ecology* 34:447–468.

Lykken, D. T. 1968. "Statistical significane in psychological reseach.". *Psychological Bulletin* 70:151–159.

Maxwell, S. E., and H. D. Delaney. 1990. *Designing experiments and analyzing data: A model comparison perspective*. CA: Wadsworth: Belmont.

McQuarrie, E. F. 2014. "Threats to the scientific status of experimental consumer psychology: A Darwinian perspective". *Marketing Theory* 14 (4): 477–494.

Micceri, T. 1989. "The unicorn, the normal curve, and other improbable creatures". *Psychological Bulletin* 105 (1): 156–166.

Miller, J. 2009. "What is the Probability of Replicating a Stastistically Significant Effect?". *Psychonomic Bulletin and Review* 16 (4): 617–640.

Morey, R. D., J. Rouder, J. Verhagen, and E. J. Wagenmakers. 2014. "Why hypothesis tests are essential for psychological science: a comment on Cumming (2014).". *Psychological Science* 25 (6): 1289–90.

Mulaik, S., N. Raju, and R. Harshman. 1997. "There is a time and a place for significance testsng". In *What if there were no significance tests?*, edited by L. Harlowand, S. Mulaik, and J. Steiger, 65–115. Erlbaum.

Murphy, K. R., and B. Myors. 1999. "Testing the hypothesis that treatments have negligible effects: minimum-effect tests in the general linear model". *Journal of Applied Psychology* 84 (2): 234–248.

Neuhaeuser, M. 2010. "An equivalence test based on n and p". *Journal of Modern Applied Statistical Methods* 9 (1): 304–307.

Nickerson, R. S. 2000. "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy". *Psychological Methods* 5 (2): 241–301.

Nieuwenhuis, S., B. U. Forstmann, and E. Wagenmakers. 2011. "Erroneous analyses of interactions in neuroscience: a problem of significance". *Nature Neuroscience* 14 (9): 1105–1107.

Nosek, B. A., J. R. Spies, and M. Motyl. 2012. "Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability". *Perspectives on Psychological Science* 7 (6): 615–631.

Parkhurst, D. F. 2001. "Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation". *BioScience* 51 (12): 1051–1057.

Pollard, P., and J. T. E. Richardson. 1987. "On the probability of making Type 1 errors". *Psychological Bulletin* 102:159–163.

Robinson, A. P., and R. E. Froese. 2004. "Model validation utestsequivalence tests". *Ecological Modelling* 176:349–358.

Rosenthal, R. 1979. "The file drawer problem and tolerance for null results.". *Psychological Bulletin* 86 (3): 638–641.

Rosnow, R., and R. Rosenthal. 1989. "Statistical procedures and the justification of knowledge in psychological science". *American Psychologist* 44:1246–1284.

Rouder, J. 2014. "Optional stopping: No problem for Bayesians". *Psychonomic Bulletin & Review* 21 (2): 301–308.

Rukhin, A., J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo. 2010. *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*. Number 800-22 in NIST Special Publication. National Institute of Standards and Technology.

Sargent, R. G., D. Goldsman, and T. Yaacoub. 2015. "Use of the interval statistical procedure for simulation model validation". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. Roeder, C. Macal, and M. C. Rossetti. IEEE.

Savalei, V., and E. Dunn. 2015. "Is the call to abandon p-values the red herring of the replicability crisis?". *Frontiers in Psychology* 245:1–4.

Schmidt, F., and J. Hunter. 1997. "Eight common but false objections to the discontinuation of significance testing in the analysis of research datat". In *What if there were no significance tests?*, edited by L. L. Harlow, S. A. Mulaik, and J. H. Steiger, 37–64. Erlbaum.

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.

Sedlmeier, P. 1996. "Jenseits des Signifikanztest-Rituals: Ergaenzungen und Alternativen". *Methods of Psychological Research Online* 1 (4): 41–63.

Senn, S. 2001. "Two cheers for P-values?". *Journal of Epidemiology and Biostatistics* 6 (2): 193–204.

Serlin, R. C., and D. K. Lapsely. 1985. "The good enough principle". *American Psychologist* 40 (1): 73–83.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". *Psychological Science* 22 (11): 1359–1366.

Soto, J. 1999. "Statistical testing of random number generators". In *Proceedings of the 22nd National Information Systems Security Conference*, 1–12. NIST.

Thompson, B. 2007. "Effect sizes, confidence intervals, and confidence intervals for effect sizes". *Psychology in the Schools* 44 (5): 423–432.

Troitzsch, K. 2014. "Analysing Simulation Results Statistically: Does Significance Matter?". In *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*, edited by D. Adamatti, G. Dimuro, and H. Coelho, 88–105. PA, USA: Hershey.

Tufte, E. 2001. *The visual display of quantitative information (2nd ed.)*. Cheshire, CT: Graphics Press.

Tukey, J. 1991. "The philosophy of multiple comparison". *Statistical Science* 6:100–116.

Tukey, J. W. 1960. "A survey of sampling from contaminated normal distributions". In *Contributions to Probability and Statistics*, edited by I. S. Olkin, W. Ghurye, W. Hoeffding, W. Madow, and H. Mann, 448–485. Stanford, CA.: Stanford University Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Pearson.

Wagenmakers, E. J., J. Verhagen, A. Ly, D. Matzke, H. Steingroever, J. N. Rouder, and R. D. Morey. 2015. "The Need for Bayesian Hypothesis Testing in Psychological Science". In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, edited by S. O. Lilienfeld and I. Waldman, in Press. University of Missouri Press (in Press).

Wasserstein, R. L., and N. A. Lazar. 2016. "The ASA's statement on p-values: context, process, and purpose". *The American Statistician* 0 (ja): 00–00.

White, J., A. Rassweiler, J. Samhouri, A. Stier, and C. White. 2014. "Ecologists should not use statistical significance tests to interpret simulation model results". *Oikos* 123:385–388.

Wilcox, R. R. 2012. *Introduction to Robust Estimation and Hypothesis Testing 3rd ed.* Elsevier Science Publishing.

Wilkinson, L. 1999. "Task force on statistical inference: Statistical methods in psychology journals". *American Psychologist* 54:594–604.

Yu, E., A. Sprenger, R. Thomas, and M. Dougherty. 2014. "When decision heutistics and science collide". *Psychonomic Bulletin & Review* 21 (2): 268–282.

AUTHOR BIOGRAPHY

MARKO HOFMANN is Chief Scientist at ITIS GmbH in Neubiberg, Germany, and adjunct Professor at the University of the Federal Armed Forces in Munich since 2010. Email: marko.hofmann@unibw.com