Meta-Psychology, 2025, vol 9, MP.2021.2927
https://doi.org/10.15626/MP.2021.2927
Article type: Original Article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Rickard Carlsson
Reviewed by: Aberson, C., Williams, M.N., Beribisky, N., Vanhove, J.
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF: https://doi.org/10.17605/OSF.IO/E4RDM

Overview on the Null Hypothesis Significance Test: A Systematic Review on Essay Literature on its Problems and Solutions in Present Psychological Science

Noah van Dongen[1] and Leonie van Grootel[2]
[1]Department of Psychology, University of Amsterdam, The Netherlands
[2]Rathenau Instituut, Den Haag, The Netherlands

For decades, waxing and waning, there has been an ongoing debate on the values and problems of the ubiquitously used null hypothesis significance test (NHST). With the start of the replication crisis, this debate has flared-up once again, especially in the psychology and psychological methods literature. Arguing for or against the NHST method usually takes place in essays and opinion pieces that cover some, but not all the qualities and problems of the method. The NHST literature landscape is vast, a clear overview is lacking, and participants in the debate seem to be talking past one another. To contribute to a resolution, we conducted a systematic review on essay literature concerning NHST published in psychology and psychological methods journals between 2011 and 2018. We extracted all arguments in defense of (20) and against (70) NHST, and we extracted the solutions (33) that were proposed to remedy (some of) the perceived problems of NHST. Unfiltered, these 123 items form a landscape that is prohibitively difficult to keep in one's sights. Our contribution to the resolution of the NHST debate is twofold. 1) We performed a thematic synthesis of the arguments and solutions, which carves the landscape in a framework of three zones: mild, moderate, and critical. This reduction summarizes groups of arguments and solutions, thus offering a manageable overview of NHST's qualities, problems, and solutions. 2) We provide the data on the arguments and solutions as a resource for those who will carry-on the debate and/or study the use of NHST.

Keywords: null hypothesis significance test; systematic review; essay literature; opinion literature; thematic synthesis

Introduction

Between 2011 and 2015, the social sciences, psychology in particular, entered what has become known as the replication crisis (Baker, 2016). For instance, the Open Science Collaboration (2015) attempted to replicate 100 landmark studies in the field of psychology, from which less than half yielded results similar to the original publication. Replication attempts in other academic fields showed comparable results, such as economics, (e.g., Camerer et al., 2016), cancer research (e.g., Begley & Ellis, 2012), and medicine (e.g., Prinz et al., 2011). In addition, Simmons et al. (2011) had shown how surprisingly easy it can be to increase the probability of obtaining positive results for a (n obviously) false hypothesis, using research practices that were considered conventional. According to many, one of the major culprits of this state of crisis, next to publication bias (Franco et al., 2014) and researcher degrees of freedom (Wicherts et al., 2016), was (the misuse of) the Null Hypothesis Significance Test (NHST) procedure for statistical inference (e.g., Cumming, 2014; Wagenmakers et al., 2011). NHST is the procedure that is ubiquitously used in the social sciences for the analysis of quantitative data. Roughly summarizing NHST, the test statistic of data that are produced by a study usually compared to a null hypothesis of no-effect (e.g., no difference between two groups on the investigated measure). The null hypothesis is rejected if the test statistic is sufficiently improbable according to this null hypothesis. Specifically, the null hypothesis is rejected if, under the assumption that the null hypothesis is true, the probability of observing this test statistic or one more extremely deviating from the null hypothesis is below a certain threshold (i.e., 5% probability; $p < .05$). In such a case, the null hypothesis is generally rejected in favor of an unspecified alternative hypothesis (e.g., there is a difference between the two groups).

Otherwise, the null hypothesis is retained, though one is not supposed to consider this as evidence in its favor. This seemingly simple procedure is associated

with two major problems: 1) is easily misused and 2) it possesses (hidden) deficiencies. There is abundant empirical evidence (e.g., see Engman, 2013) of social scientists[1] unwittingly misusing the method and misinterpreting its results (e.g., interpreting the probability of the data under H0 as the probability of H0 given the data). With regard to the second problem, deficiencies and flaws of NHST that are independent of misuse have been brought into the limelight (e.g., ease of rejecting H0 with a large enough sample when it is not perfectly true; Wagenmakers, 2007).

These issues, in combination with publication bias (i.e., favoring positive results for publication; Sutton, 2009) and a publish-or- perish incentive structure (i.e., academic careers depending on the number of publications and the journals in which they are published; Szucs and Ioannidis, 2017) are expected to be at the root of the low replicability rate of social science research. As a response to the first mentioned problem, NHST's misuse, several proposals have been made in the literature. People have called for and are starting to implement more transparency (e.g., open data and methods; Freese & King, 2018; Munafò, 2016), and design and analysis specification prior to data collection (i.e., preregistration; Lindsay et al., 2016; van 't Veer & Giner-Sorolla, 2016), which makes misuse of NHST, like additional sampling, more difficult. For instance, registered reports have, at the moment of writing, been adopted by more than 250 journals (Center for Open Science, 2020). Roughly speaking, a registered report is a study design and analysis plan that is submitted for peer review at a journal (that accepts registered reports) prior to the data collection and accepted on its methodological merits and general theoretical relevance alone. Early meta-scientific results on the effectiveness of registered reports are promising (Scheel et al., 2020).

In response to the second mentioned problem, NHST's deficiencies, researchers and statisticians have called for reform in statistical methodology. Some have proposed that, if we were to retain NHST, we should at least lower the threshold for statistical significance (Benjamin et al., 2018), which would reduce the problem of inflated false-positive rates to some extent. This view is opposed by those that think that there should not be a default threshold and that the critical value should be set and justified by the study design, the research question, etc. (Lakens et al., 2018).

Then there are others who want to abandon the statistical significance threshold altogether (McShane et al., 2019). In this case, the advice is to refrain from hypothesis tests and use, for instance, only estimation methods instead (e.g., Cumming, 2014). This perspective is, in turn, opposed by those who see the merits of hypothesis testing and binary decision-making in the social sciences (e.g., Morey et al., 2014). Instead of estimation, they propose a Bayesian version of hypothesis testing that is supposed to be less prone to error and misuse (e.g., Wagenmakers et al., 2017). All these groups agree that the current practice involving NHST is detrimental to science. However, there is a worrisome absence of opinions and suggestions of these groups converging on a semblance of consensus about how these problems should be addressed.

The literature on the problems of the significance test goes back almost a century. Wilson (1923) criticized the use of significance tests even before Fisher developed the likelihood and *p*-values. Berkson (1942) argued against the use of significance tests as evidence and expounded on the problems of having no specific alternative hypothesis. Hogben (1956) highlighted the problem of testing a null hypothesis without setting the sample size in advance. Rozeboom (1960) argued against NHST as a procedure that is of value to science, because of the absence of inverse probabilities. Bakan (1966), Greenwald (1975) and Rosenthal (1979) discussed the inherent bias against the null hypothesis in the application of NHST and the deleterious consequences this has on reported research. Berger and Sellke (1987) demonstrated the problems of interpreting *p*-values as a measure of evidence against the null hypothesis. Cohen (1994) discussed many misapplications of NHST, among them the misinterpretation of *p*-values as the probability that the null hypothesis is false, the misinterpretation that one minus the *p*-value is the probability of successful replication, and the mistaken assumption that if one rejects the null hypothesis based on a *p*-value below 0.05 one thereby affirms the theory that led to the test. And, in 2000, Nickerson published a non-systematic review that discussed the problems mentioned above and several others (e.g., the null hypothesis can almost never be perfectly true and the interpretation of *p*-values as effect sizes).

Although much has been published about the problems with NHST and its application, the method remained ubiquitous in the social sciences. The current attempt at methodological reformation is also not the first time that scientists have tried to change the NHST system. For instance, around 2000 there was a serious collective attempt to change the way in which scientists analyze data and report results (e.g., Cohen, 1994; Gigerenzer, 2004; Meehl, 1990, 1992). Meehl (1990, p.193) put it succinctly when he wrote:

"Let me say as loudly and clearly as possi-

[1] For brevity, we will drop 'social' and refer to social scientist as scientists

> ble that what we critics of weak significance testing are advocating is not some sort of minor statistical refinement (e.g., one-tailed or two-tailed test? unbiased or maximum likelihood statistics? pooling higher order uninterpretable and marginal interactions into the residual?). It is not a reform of significance testing as currently practiced in soft psychology. We are making a more heretical point than any of these: We are attacking the whole tradition of null-hypothesis refutation as a way of appraising theories. The argument is intended to be revolutionary, not reformist.

Looking back, these events did however not inspire the intended methodological revolutions. This could be because papers about (how to solve) NHST's problems are focused on only a small number of issues and are contraindicated in other papers by authors who disagree, instead of forming a coherent structure that provides a clear plan of attack. For instance, Cohen (1994) claimed that we, scientists, are interested in the probability that the hypothesis is true and should thus use (something akin to) Bayesian statistics. This was then contradicted in responses to Cohen's paper (e.g., Baril & Cannon, 1995; Frick, 1995; McGraw, 1995; Parker, 1995). In turn, Cohen tempered his claim in a rejoinder (Cohen, 1995). As it appears, many authors agree on the fact that there is a problem to be addressed concerning NHST, but most differ in which deficiencies and misuses of the method they address. The papers seem to talk past each other, each vying for attention, and in doing so, inhibiting improvement of the practice of statistical inference. A clear overview of the perspectives on NHST is absent and without such an overview - a map of NHST's qualities, problems, its solutions - it will be nigh impossible to plot a course of action away from the replication crisis and towards methodological health. To this end, a framework is required in which the different perspectives on NHST can be placed side-by-side, compared, and evaluated. Such a framework will allow us to assess the veracity of each argument for or against NHST and assess the potential consequences of proposed methodological reform prior to implementation.

With the systematic review reported in this paper, we aim to provide the foundations of such a framework. By collecting and structuring the arguments for and against NHST put forth in academic publications that were published in psychology over the last decade, we provide an overview of the perspectives put forth during the current attempt at methodological reformation. Furthermore, the overview of the discussion and the dataset that is established through this research can provide valuable information and data for other researchers to continue with.

Method

We reviewed opinion literature describing the misconceptions and limitations of the Null Hypothesis Significance Test (NHST). We synthesized essays and opinion pieces, because we are interested in the arguments researchers use to make their point for or against NHST and the solutions they provide. To our knowledge, at the moment of writing, there are no specific guidelines for the synthesis of opinion literature yet. Since our review question is mainly exploratory and open, we made use of existing guidelines for reviews of qualitative and mixed methods studies.

To ensure transparency, extension, and reproducibility, we adhere to the ENTREQ transparency statement for qualitative research (Tong et al., 2012). The ENTREQ was developed for qualitative reviews of medical research to ensure that all relevant information is explicitly stated in the research report. To aid this endeavor, Table 1. provides the list of applicable ENTREQ items and in which sections the information is given that pertains to that item. Five out of 21 ENTREQ items were not addressed. One item (ENTREQ: 18.) is concerned with study comparison, which is not applicable to our review. Four items (ENTREQ: 10. - 13.) are concerned with the quality appraisal of the data, which is absent from our review. We considered the checklist for text and opinion tools from the Joanna Briggs Institute (2017), but we preferred to focus on the quality of the individual arguments that make up the opinion that is to be evaluated which this tool does not provide. Furthermore, using the tool in spite of this deficiency could have led to the exclusion of papers based on their overall opinion, whereas articles with 'illogical' opinions could still contain arguments with interesting information for our analysis (e.g., Carroll et al., 2012). The rest of this section contains the formulation of the review question, the search strategy, the selection of studies, the data extraction, and the data analysis plan.

Review question

We used the acronym SPIDER (Cooke et al., 2012) to aid us in the focus on a review question. The SPIDER acronym allows the reviewer to specify the sample, phenomena of interest design, evaluation, and research type. We specified the components of the SPIDER acronym as follows:

- Sample: Opinion statements (conclusions) supported by arguments for or against the Null Hy-

Figure 1

Selection process of essay and opinion pieces pertaining to the qualities and problems of NHST.

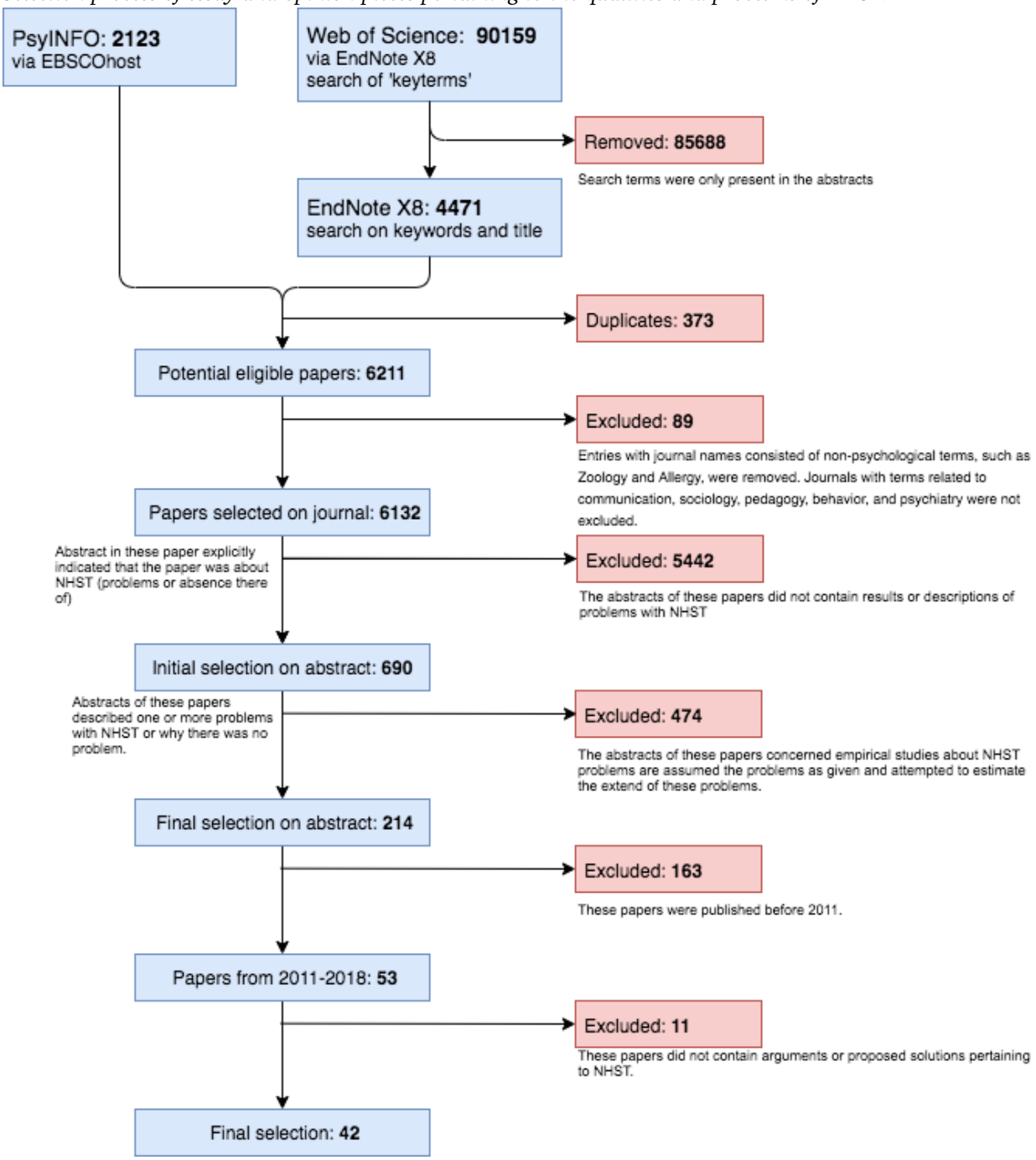


Table 1

ENTREQ items adhered to and where they can be found in the text.

No.	Item	Location
1	Aim	Section 2.1.
2	Synthesis methodology	Section 2.5.
3	Approach to searching	Section 2.2.
4	Inclusion criteria	Section 2.2., Appendix S
5	Data sources	Section 2.2., Appendix S
6	Electronic search strategy	Section 2.2., Appendix S
7	Study screening methods	Section 2.4., Appendix S
8	Study characteristics	Section 3.1.
9	Study selection results	Section 2.2.
14	Data extraction	Appendix E
15	Software	Section 2.5.
16	Number of reviewers	Section 2.4., Section 2.5.
17	Coding	Section 2.5.
19	Derivation of themes	Section 2.5.
20	Quotations	Section 3.2.
21	Synthesis output	Section 3.2.

pothesis Significance Test method and proposed solutions to problems, which in turn are based on evidence in the form of literature references, examples, mathematical proofs, and/or data simulations.

- Phenomena of Interest: deficiencies of NHST, misconceptions about NHST, defenses of NHST, solutions to problems pertaining to NHST, and overall conclusions about NHST according to the authors of the papers. NHST deficiencies: Intrinsic properties of the method that are considered undesirable, limiting or pernicious (e.g., *p*-value is not an absolute measure of evidence, but relative to the sample size and effect-size of the population). NHST misconceptions: Misconceptions about the results of the Null Hypothesis Significance Test are specific misinterpretations of the results produced by the NHST method that are linked to (elements of) the method (e.g., interpreting a *p*-value as the probability of the null hypothesis on the observed data instead of the probability of the data on the assumption that the null hypothesis is true). NHST defenses: Properties of NHST that speak to its qualities or counterarguments to the misconceptions and deficiencies. Solutions to NHST: Solutions that are purported to mitigate or remedy the mentioned problem. Examples of such solutions are: a) alternative methods of statistical inference that supposed to suffer less from limitations and potential for misconception than the NHST procedure; b) safeguards or policy changes that mitigate or remedy the purported problems; c) changes to the NHST method mitigate or remedy the purported problems. Conclusions: The particular conclusions about, perspectives on, or positions to NHST the writers have that follow from / are built on the arguments they use in their paper.
- Design: Discussions / essay on the phenomena of interest using arguments supported by evidence.
- Evaluation: Identification and description of perspectives and opinions on the NHST.
- Research type: Essays and opinion pieces (i.e., non-empirical scientific publications) in psychological science and the accompanying methodological literature published between 2011 and 2018.

This boils down to the following two general review questions:

What are the deficiencies, misconceptions, and defenses of NHST?

What are the proposed solutions to problems pertaining to NHST?

Search Strategy

Although this is a novel type of systematic review, we tried to adhere as much as possible to reporting standards already in existence. We adopted the STARLITE reporting standard for literature searches (Booth,

2006). STARLITE is a reporting standard developed for qualitative systematic reviews from a survey of existing reviews with the aim to improve the quality and completeness of systematic review reports. We explicate the items that make up the STARLITE mnemonic:

- Sampling strategy: We used purposive sampling. Specifically, we restricted our search to papers published in psychology, psychological methods or statistics literature.
- Type of study: As indicated in section 2.1, we limited this review to essays and opinion pieces. We chose these types of papers because this is where the pros, cons, and solutions of NHST are presented.
- Approaches: We used electronic subject searches. No other approaches were used, because we expected that all published papers on the subject are reachable through electronic searches on online databases.
- Range of years: We included papers published from 1 January 2011 up to 31 December 2018. We choose this restriction because 2011 is recognized as the year the replication crisis started and psychology is the field in which it takes place or, at least, is most engaged with investigating and remedying the situation.[2] We selected 1 January as the starting date because Bem's paper Feeling the Future on precognition, which triggered the replication crisis, was published in January 2011 and that received some media coverage at the end of 2010 (e.g., Lehrer, 2010). We performed the search in January 2019. We decided to focus on the last decade and the replication crisis, because we consider these publications to contain perspectives held by people who are active in the current attempt at methodological reformation.
- Limits: The functional limits that were applied concern English language and publication in peer reviewed scientific journals. The literature was limited to peer reviewed journal publications because an adequate quality assessment instrument is lacking for essay literature and peer review offers some quality control (e.g., Ma et al., 2013). Inclusion and exclusion: Papers were included in they were about the Null Hypothesis Significance Test inference procedure (i.e., the correct Type of study); either wholesale or parts of the procedure. Specifically, papers were included if they were about the concepts '*p*-value', 'statistical significance', 'significance tests', or 'null hypothesis'. Papers were excluded if they were about empirical research and the concepts were used in the paper, but the papers were not about these concepts (e.g., "the results were statistically significant at the $p<0.05$ level"); or if the papers were about another statistical procedure that has similar concepts (e.g., Bayesian hypothesis test) and the abstract did not indicate a comparison to NHST. Figure 1 depicts more specific information about the reasons for exclusion of papers.
- Terms used: We used the following search terms: 'null hypothesis', 'hypothesis test', 'statistical significance', 'significance test', '*p*-value', and 'p value'. We used these terms, because these are the concepts most commonly associated with NHST. All search terms with syntax, operators, and exclusion criteria are provided in Appendix S.
- Electronic sources: We used two databases for the literature search, PsycInfo (via EBSCOhost) and Web of Science Core (via EndNote 8). We limited our search to these two databases, because they cover the width of psychological literature.

The search was conducted by NVD and was checked for errors and inconsistencies by [LVG]. Identified ambiguities and inconsistencies (e.g., does a journal meet our exclusion criteria) were discussed and resolved on mutual agreement. The schematic representation of the selection process is presented in Figure 1. The initial search yielded 6221 potential papers and the final selection consisted of 42 papers.

Data Extraction

We extracted descriptive information on the articles and descriptive information concerning the authors' opinion on NHST. The descriptive information on the article consisted of the name of the (first) author, year of publication and journal, and whether or not the article was part of a special issue. Furthermore, we extracted the purpose of the article if explicitly stated by the author(s) in the abstract or introduction; the readability for non-expert (i.e., absence of advanced mathematics and assumed technical know-how in the argumentation); the applicability of solutions (i.e., solutions

[2] We acknowledge that there are several prior papers that contributed to the commencement of what is now called the replication crisis. However, Bem's Feeling the Future paper can be considered the proverbial straw that broke the camel's back. It elicited a flurry of comments and was a major motivator for the first large replication project in psychological science, which started in November 2011 (Open Science Collaboration, 2015).

do not require costly statistical software or mathematical experience); and the conclusion, if explicitly stated, in the last section of the article. Descriptive information concerning the authors' opinion on NHST consists of arguments and solutions related to NHST was extracted from all parts of the papers. Arguments and solutions were identified in terms of explicit claims pertaining to the phenomena of interest (as specified in Section 2.1). Specifically, a description in one or more sentences on their position (or the views/position of others they endorse) with respect to, for instance, a particular deficiency of NHST. An example of data extraction of an argument from one of the papers is as follows. On the first page, Trafimow (2013) states: "*p*-values generally are inaccurate estimators of probabilities of null hypotheses." This claim indicates NHST's deficiency that its outcome measure, the *p*-value, does not provide an accurate estimation of the probability of the null hypothesis. It was therefore interpreted as a deficiency of NHST. Note that all extracted arguments, solutions, and conclusions are claims made by the authors of the papers and are not necessarily valid.[3]

Evidence used to support these claims was also extracted. Empirical research, examples, other literature, simulations, and/or mathematical proofs were extracted as evidence if they were referenced for that claim. For example, on page 13, Cumming (2014) makes the claim: "[w]hatever the N, a [*p*-value] gives only extremely vague information about replication (Cumming, 2008)." In this case, "(Cumming, 2008)" is used as support for this claim. We interpreted this information as evidence of the type: reference to other literature. The extraction form, provided as Appendix E, gives further details concerning what was extracted from the indicators that were for their classification. The data were extracted by [NVD] and the data extraction procedure was checked for errors and inconsistencies by [LVG]. Identified ambiguities and inconsistencies, such how to categorize a certain argument, were discussed and resolved on mutual agreement.

Data Analysis

The purpose of our systematic review is to provide an overview of the qualities, deficiencies, and misconceptions of, and solutions to NHST. The collection of extracted arguments, arguments, proposed solutions, conclusions, and their supporting evidence provide this overview, but it is too vast to interpret without some kind of focus or filter. To this end, we conducted an initial qualitative thematic synthesis (Thomas & Harden, 2008) in which we identified clusters of arguments and solutions described in the articles.

Thematic synthesis

The thematic synthesis contains four steps: 1) identification of claims in the papers, 2) development of descriptive themes, 3) the development of analytical themes within the phenomena of interest, and 4) development of overarching analytical clusters. These four steps were as follows. First, [NVD] identified pertinent text fragments (i.e., the claims pertaining to arguments, solutions, or conclusions) from all sections of the articles (top left of the ladder in Figure 2). Second, [NVD] developed descriptive themes (Thomas & Harden, 2008) by clustering fragments into groups: using the software program EPPI-reviewer 4.11.4.0 (Thomas et al., 2020). [NVD] and [LVG] discussed the descriptive themes upon agreement. Text fragments were grouped in descriptive themes based on interpreted overlap in content, meaning, or intention.[4] These descriptive themes were the individual arguments concerning NHST's deficiencies, misconceptions, and qualities, and solutions and conclusions (second step in Figure 2). Third, from the 137 descriptive themes, analytical themes were developed Thomas and Harden (2008). Analytical themes go beyond the information from the primary studies in order to address our review question directly. During this step, the analysis was done separately for each phenomenon of interest as specified in the SPIDER research question: Deficiencies, misconceptions, NHST defense arguments, solutions, and conclusions. The descriptive themes were first grouped in analytical subthemes, based on similarities in their semantic content. These analytical subthemes were then grouped under overarching analytical themes within each phenomenon of interest (step three and four in the ladder of Figure 2). Four, overarching analytical clusters[5] were developed from the analytical themes identified within the phenomena of interest. These overarching clusters form semantically coherent zones within the landscape of NHST, containing its qualities, problems and the solution to these problems.

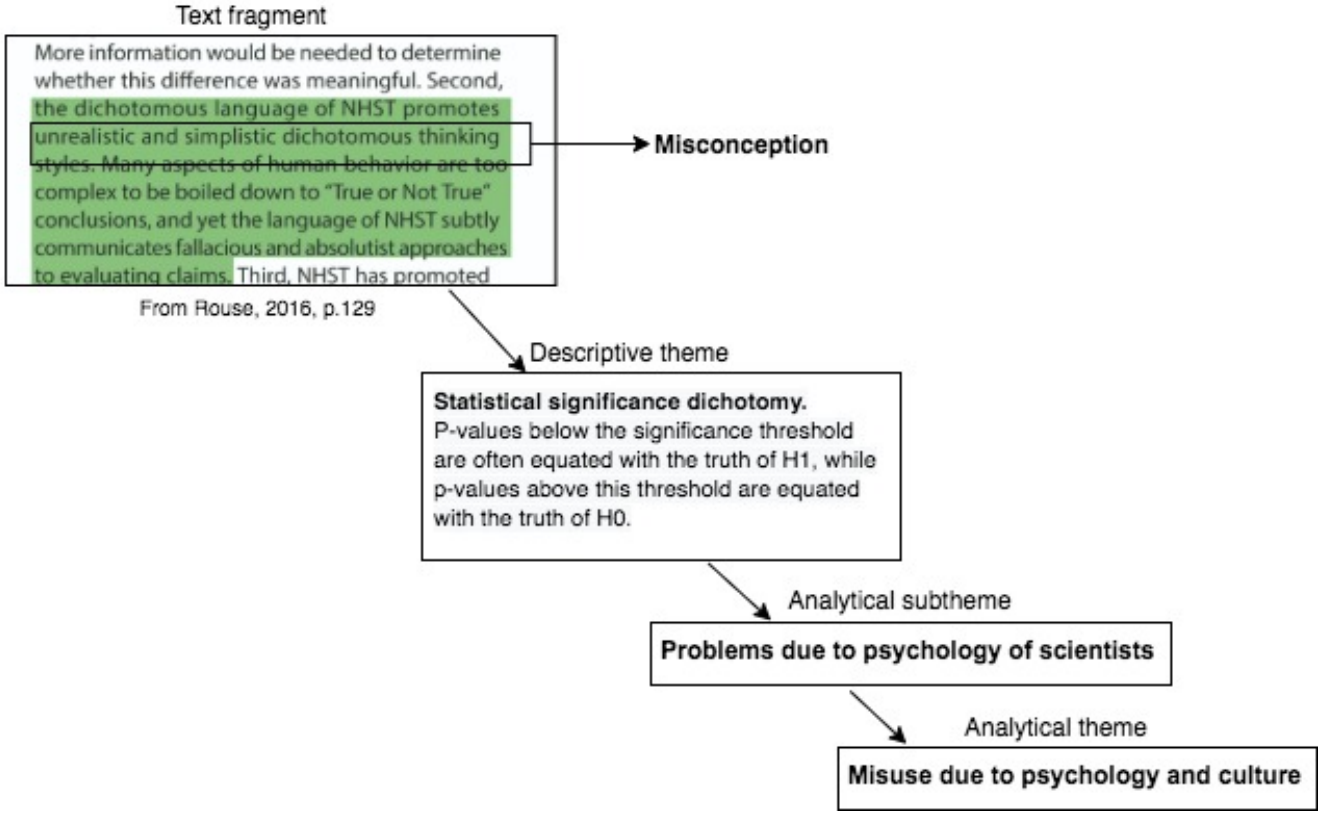
[3] We decided to include all arguments without any judgment on their validity, because we did not possess a systematic approach to evaluate the arguments, neither in isolation nor in relation to each other. In addition, (in)validity of arguments seems to be highly dependent on one's perspective on scientific inference / philosophy of science, making a categorical judgment impossible.

[4] Note that this is a slight deviation from Thomas and Harden (2008). We adopted this deviation to make the method amenable to the unorthodox nature of our systematic review.

[5] We adopted this term to add an additional layer in the hierarchy that provides some conceptual distinction.

Figure 2

Example of the analysis hierarchy within a phenomenon of interest. A text fragment is first labeled with a descriptive theme. These descriptive themes are then clustered in analytical subthemes based on their similarity in semantic content. These were then clustered into higher-order analytical themes within each phenomenon of interest.



Results

The results section describes the descriptive information concerning the included papers and the themes generated in the thematic synthesis.

Thematic Synthesis

A total of 42 articles were included in the analysis. Of these, 34 were regular articles, two editorials and six commentaries. Eleven articles were part of a special issue; of which eight were published in "Perspectives on the Use of Null Hypothesis Statistical Testing (NHST)" a three-part special issue edited by Educational and Psychological Measurement. Most papers (12 out of 42) were published in 2017 and only a few papers were published in 2012, 2014, and 2018 (respectively, 1, 3, and 2 out of 42). All but one of the papers were considered readable. The exception was a paper published in Statistica Sinica (Martin & Liu, 2014) and required the understanding of mathematical notation, the stated mathematical theorems and their proofs. Out of 42 articles, 29 offered one or more potential solutions to the problems pertaining to NHST. In only three out of these 29 cases, no relevant information was given on how to implement the suggested solution(s). A complete overview of the paper characteristics can be found in Table 2.

Arguments, solutions, conclusions, and evidence

We identified a total of 90 arguments, 33 proposed solutions and 14 conclusions across the 42 articles that were included. These 137 items provide a first step in answering the review question: they describe the NHST landscape. Concerning the arguments, we identified 39 deficiencies of NHST, 31 arguments about how NHST is misinterpreted and misused, and 20 on the defense of NHST. A rough overview of the distribution of arguments and solutions across the papers is presented in Table 3. The complete lists of these arguments, solutions, and conclusions, their descriptions, and from which pa-

pers they originate can be found in Appendix A. The complete dataset that contains further details can be accessed https://osf.io/bktfx/). As is shown in Table 2., the range in the number of items per paper is large (i.e., from 1 to 30 per paper). One paper contained as many as 11 deficiencies. Another paper contained as many as 14 misconceptions. The maximum number of NHST defenses was 8. The highest number of solutions was 11. Not a single paper had the highest count on more than a single category.

In total, we identified 508 unique pieces of evidence that were referenced in the 42 papers to support the 123 arguments and solutions. 15 were published before 1950; 139 were published between 1950 and 1999; 181 were published between 2000 and 2010; and 173 were published between 2011 and 2017 (see Figure 3. for the distribution of publication dates). The body of evidence is diverse. For instance, some references that were used as evidence for the arguments in the papers came from journals such as Journal of Wildlife Management, The Journal of Experimental Education, Philosophy of Science, Political Analysis, Biological Psychiatry, Seminars in Hematology, Ecology, and Biology of Blood and Marrow Transplantation. A detailed analysis and interpretation of the evidence extracted from the papers is beyond the scope of our review. We can note that papers that can be considered as classics, like Berkson (1942), Bakan (1966), Rosenthal (1979), Cohen (1994), and Nickerson (2000) are on several occasions used as evidence to support arguments, solutions, and conclusions. For instance, Cohen (1994) was used 22 times to support arguments and conclusions across the 42 papers. See the online supplements for further details on the extracted evidence, https://osf.io/ayf56/.

Thematic Synthesis Results

The opinions on NHST as presented in the literature were analyzed using thematic synthesis. The individual arguments, solutions, and conclusions were used as the unit of analysis. We generated three overarching analytical clusters from the analysis that can be identified as three zones in the (problems and solutions concerning) NHST landscape: 1) the mild problems zone 2) the moderate problems zone and 3) the critical problems zone. For all three zones, we describe the identifying element(s) and we specify the associated arguments, solutions and conclusions, and provide quotes from the papers to support our synthesis. The terms in italics represent the analytical subthemes generated during the analysis (see Section 2.5.). See Figures 4, 5, and 6 for an overview of the mild, moderate, and hazardous zone of the NHST landscape respectively.

The Mild Zone

The mild problem zone of NHST (see Figure 4) can be labeled as: NHST itself is not the problem. The problems with statistical inference, such as inflated false-positive rates and biased results, have to do with misconceptions and faulty practices due to the psychology of the researchers and the research culture, not with the method that is being used. Specifically, these problems result from psychological heuristics that lead to misunderstanding, the tendency to dichotomous reasoning, and bad research practices that are considered acceptable by the community. This means that the problems have not so much to do with the method per se, but with peripheral aspects, like the language of NHST in relation to human reasoning and behavior. For instance, Rouse (2016, p.129) writes

> ... the dichotomous language of NHST promotes unrealistic and simplistic dichotomous thinking styles. Many aspects of human behavior are too complex to be boiled down to “True or Not True” conclusions, and yet the language of NHST subtly communicates fallacious and absolutist approaches to evaluating claims.

Within this zone, solutions are geared towards leaving the method as it is, though adding supplements and safeguards to NHST use, that mitigate the psychological and cultural problems.

These solutions consist of policy changes (e.g., in education, funding, publication), focus on proper use of NHST (e.g., high enough power, adherence to strict guidelines), and additions to NHST, such as procedure verification, meta-analyses, and direct replication. For instance, Rouse (2016, p.132) writes:

> "As the research community is examining and debating the most appropriate use of inferential statistics, we recommend that researchers continue to use the tools of NHST, but to use these tools carefully and effectively."

This zone also contains the analytical themes of defenses of NHST and its use. These themes are 1) other methods are just as bad or worse: Bayesian hypothesis tests and estimation methods either come with additional and more pernicious problems (for the former) or are based on the same principles and thus suffer from the same problems (for the latter). 2) NHST does what it is supposed to do: the shortcomings, such as sampling plan affecting results, are actually qualities and what NHST is supposed to do, controlling error rates, it does

well. 3) NHST is good enough for what scientists need and can do/have: NHST is actually quite intuitive to understand or at least not more misunderstood than other methods and can be unproblematically operated when certain precautions are taken (e.g., sufficient power and replication studies). For instance, Miller (2017, p.664) writes:

> To the contrary, the heart of NHST is a simple, intuitive, and familiar "common sense" logic that most people routinely use when they are trying to decide whether something they observe might have happened by coincidence (a.k.a., "randomly," "by accident," or "by chance").

This zone can be seen containing the milder conclusions concerning NHST, such as the analytical subthemes use NHST with care and proper practice, use complete reporting, and improve general practice. In other words, from within this zone, there is nothing actually wrong with NHST. Specifically, the method works as it is supposed to work when used correctly. Policy and education changes are all that is required to remedy the existing problems.

The Moderate Zone

The second zone (see Figure 5.) within the NHST landscape contains the proposals for moderate, but not fundamental, changes in light of serious misuse of NHST and mild shortcomings of the methods. Within the borders of this zone, the general conclusion is that researchers should stick with frequentist hypothesis testing, just in a different form than NHST. This differs from the mild zone in the sense that the NHST approach, with its singular reliance on *p*-values to accept unspecified alternative hypotheses, should be dropped in favor of different approaches to hypothesis testing, though are still based on the frequentist philosophy of probability (e.g., interval-null hypothesis testing). Scientists misunderstand how NHST works which leads to wrongfully interpreted results. This misunderstanding is caused by deficiencies concerning the input for the analysis and the output. To illustrate, within this zone, the use of NHST is problematic because, for instance, if the sample size is large enough, the null hypothesis will always be rejected as a result if it is not perfectly true (e.g., "Berkson (1938) was the first to notice dependence of significance testing on the sample size. He objected that it is possible to obtain a statistically significant chi-square test merely by increasing sample size" Rao & Lovric, 2016, p. 13) and, as LeBel and Peters (2011, p. 374) states, it outputs ambiguous evidence:

> Thus, although it is well known that negative (null) results are ambiguous and difficult to interpret, exclusive reliance on NHST makes positive results equally ambiguous, because they can be explained by flaws in the way NHST is implemented rather than by a more theoretically interesting mechanism (Meehl, 1967).

Although these are important criticisms on NHST as it is currently practiced, not all frequentist hypothesis testing of statistical significance needs to be sensitive to these deficiencies. As another example, the misconceptions, such as that statistical significance comes in degrees (e.g., '$0.05<p<0.1$' marginally significant, '$p<0.001$' highly significant and that high *p*-values indicate that the null hypothesis is true, can be addressed by reconceptualizing several components of NHST (i.e., what error probabilities are); as Haig (2017, p. 496) writes about adopting Error statistics theory of Mayo (1999):

> In her initial formulation of the error-statistical philosophy, Mayo (1999) modified, and built upon, the classical Neyman–Pearsonian approach to ToSS [tests of statistical significance]. However, in later publications with Spanos (e.g., Mayo & Spanos, 2011), and in writings with David Cox (Cox & Mayo, 2010; Mayo & Cox, 2010), her error-statistical approach has come to represent a coherent blend of many elements, including both Neyman–Pearsonian and Fisherian thinking. For Fisher, reasoning about p values is based on post-data, or after-trial, consideration of probabilities, whereas Neyman and Pearson's Type I and Type II errors are based on pre-data, or before-trial, error probabilities. The error-statistical approach assigns each a proper role that serves as an important complement to the other (Mayo & Spanos, 2011; Spanos, 2010). Thus, the error-statistical approach partially resurrects and combines, in a coherent way, elements of two perspectives that have been widely considered to be incompatible. In the post-data element of this union, reasoning takes the form of severe testing, a notion to which I now turn.

From which Haig (2017, p. 504) concludes:

> In more than 50 years of preoccupation with these tests, psychology has concentrated its gaze on teaching, using, and criticizing NHST in its muddled hybrid form. It

> is high time for the discipline to bring itself up-to-date with best thinking on the topic, and employ sound versions of ToSS in its research.

In addition, deficiencies, such as claims that the null hypothesis is always false or that NHST overestimates the evidence against the null hypothesis, are remedied when minor structural changes are made. These changes are, for instance, using context-dependent Type I and Type II errors (i.e., incorporating the costs of making a Type I and/or Type II error for the particular test at hand); interval null hypothesis testing (i.e., using a range of values around the null effect instead of a point null hypothesis); and using a minimum effect-size of interest null hypothesis (i.e., using as null hypothesis the smallest effect-size that would be of interest for the particular research question). As an example, Rao and Lovric (2016, p. 15) writes:

> So, what should we do? This article is an initial contribution to making a paradigm shift in testing statistical hypotheses. Instead of testing highly problematic and almost surely false point null hypotheses, as a natural replacement, test a negligible null hypothesis.

This collection of suggested solutions, and this zone in general, is succinctly expressed by Baird and Duerr (2016, p.113):

> By implementing the concepts from these sources, both traditional and contemporary, researchers are engaged in what could be described as "context-driven NHST" or CD-NHST. Instead of being driven by convention, which may or may not have much relevance, CD-NHST places the researcher in the driver's seat of inference.

Consequently, the deficiencies of NHST do require adjustments to the method, but the underlying assumption of falsification and controlling error rates remains intact.

The Hazardous Zone

The hazardous zone (see Figure 6) calls for fundamental changes in the way the social sciences practice statistical inference. This is because the foundational problems of NHST, such as its insensitivity to the specifics of the alternative hypothesis, are severe. If the alternative hypothesis is what the researcher is actually interested in, then a method is required that takes the characteristics of this hypothesis into account when evaluating the data. As Dienes and Mclatchie (2018, p. 214) write:

> Bayes factors are sensitive to how vague or precise the theory is; *P*-values are not. But, normatively, precise theories should be favored over vague ones when data appear within the predicted range.

In addition, NHST is incapable of providing that which researchers misconceive it to provide, such as interpreting *p*-values as the probability of the null hypothesis, the replication probability of the results, the realized Type I error probability, the methodological quality of the study, or that the results are due to chance. This is because the method only provides, under the assumption that the null hypothesis is true, the probability of observing the data at hand or data deviating more strongly from the null hypothesis. As an example, Schneider (2015, p. 423) writes:

> Another variant is to believe that p indicates the probability that a result is due to chance alone (i.e., sampling error). This is also not so, as p values are calculated on the assumption that H0 is true, this is the assumed 'chance model', so the probability that chance is the only explanation of the result is already taken to be 1. It is therefore illogical to view p as somehow measuring the probability of chance (Carver 1978).

Other methods might not suffer these deficiencies, making their result interpretations correct that would have been wrong for NHST results. For example, estimation methods and other types of hypothesis tests (e.g., Bayesian hypothesis tests) do not have a dichotomous decision rule (i.e., reject H0 versus retain H0) nor do they provide the probability of the data under the null hypothesis. They are said not to suffer from the problems mentioned above, because these methods are not about testing a single null hypothesis and evidence comes in degrees. As Schneider (2015, p. 427) writes:

> First of all, there are inferential alternatives, which contrary to NHST do in fact assess the degree of support that data provide for hypotheses, [...] [such as] likelihood inference (e.g., Royall 1997).

Table 2

Overview of the papers and their characteristics. The details on the special issues and the subjects of the commentaries can be found in the online dataset. The cells in the 'Useful' column are empty when the paper in question does not contain a solution.

First Author	Year	Journal	Context	Special Issue	Readable	Solutions
Gelman	2013	Journal of Mathematical Psychology	Commentary	Yes	Yes	No
Rouse	2016	Psi Chi Journal of Psychological Research	Editorial	No	Yes	Yes
Gelman	2018	Personality and Social Psychology Bulletin	Article	No	Yes	Yes
Dienes	2018	Psychonomic Bulletin and Review	Article	Yes	Yes	Yes
Campitelli	2017	Educational and Psychological Measurement	Article	Yes	Yes	Yes
Haig	2017	Educational and Psychological Measurement	Article	Yes	Yes	Yes
Haggstrom	2017	Educational and Psychological Measurement	Article	Yes	Yes	No
Schneider	2015	Scientometrics	Article	No	Yes	Yes
Wilcox	2017	Educational and Psychological Measurement	Article	Yes	Yes	Yes
Trafimow	2015	Basic and Applied Social Psychology	Editorial	No	Yes	No
Chen	2017	NeuroImage	Article	No	Yes	Yes
Trafimow	2013	Frontiers in Psychology	Commentary	No	Yes	No
Marsman	2017	Educational and Psychological Measurement	Article	Yes	Yes	No
Lu	2015	Shanghai Archives of Psychiatry	Article	No	Yes	Yes
Hupé	2015	Frontiers in Neuroscience	Article	No	Yes	Yes
Laber	2017	Journal of the American Statistical Association	Commentary	No	Yes	No
Rao	2016	Journal of Modern Applied Statistical Methods	Article	No	Yes	Yes
Goddard	2015	METODE Science Studies Journal	Article	No	Yes	Yes
Cumming	2014	Psychological Science	Article	No	Yes	Yes
Cumming	2013	Australian Psychologist	Article	No	Yes	Yes
Klugkist	2011	International Journal of Behavioral Development	Article	No	Yes	Yes
Garcia-Perez	2017	Educational and Psychological Measurement	Article	Yes	Yes	Yes
van Helden	2016	Nature Methods	Commentary	No	Yes	No
Perezgonzalez	2014	Theory & Psychology	Article	No	Yes	Yes
Citrome	2011	Bulletin of Clinical Psychopharmacology	Article	No	Yes	Yes
Konijn	2015	Communication Methods and Measures	Article	Yes	Yes	Yes
Hasley	2015	Nature Methods	Article	No	Yes	Yes

Continuation of Table 3

First Author	Year	Journal	Context	Special Issue	Readable	Solutions
Chang	2017	Educational and Psychological Measurement	Article	Yes	Yes	No
LeBel	2011	Review of General Psychology	Article	No	Yes	Yes
Miller	2017	Educational and Psychological Measurement	Article	Yes	Yes	Yes
Lambdin	2012	Theory & Psychology	Article	No	Yes	No
Hofmann	2011	Psychotherapy	Commentary	No	Yes	No
Perezgonzalez	2015	Frontiers in Psychology	Article	No	Yes	Yes
Szucs	2017	Frontiers in Human Neuroscience	Article	No	Yes	Yes
Savalei	2015	Frontiers in Psychology	Article	No	Yes	Yes
Bradley	2016	Psychological Reports	Article	No	Yes	No
Baird	2016	Journal of Modern Applied Statistical Methods	Article	No	Yes	Yes
Garamszegi	2017	Trends in Neuroscience and Education	Commentary	No	Yes	Yes
Martin	2014	Statistica Sinica	Article	No	No	Yes
Hubbard	2011	Journal of Applied Statistics	Article	No	Yes	No
Johanson	2011	Scandinavian Journal of Psychology	Article	No	Yes	Yes
Engman	2013	Quality & Quantity: International Journal of Methodology	Article	No	Yes	No

Table 3

Overview of the number of arguments and solutions per paper. See Appendix A for details on which argument or solution is present in which paper and in how many papers each item is present.

First Author	Year	Deficiencies	Misconceptions	NHST defenses	Solutions	Total
Gelman	2013	3	0	0	0	3
Rouse	2016	0	3	4	0	7
Gelman	2018	1	2	0	5	8
Dienes	2018	3	0	0	1	4
Campitelli	2017	0	0	0	1	1
Haig	2017	1	3	0	2	6
Haggstrom	2017	0	8	0	0	8
Schneider	2015	11	8	0	10	29
Wilcox	2017	0	0	0	1	1
Trafimow	2016	0	3	0	0	3
Chen	2017	2	3	2	1	8
Trafimow	2013	4	0	0	0	4
Marsman	2017	2	0	0	0	2
Lu	2015	1	1	0	1	3
Hupé	2015	6	1	0	3	10
Laber	2017	0	0	1	0	1
Rao	2016	2	0	0	1	3
Goddard	2015	1	0	0	1	2
Cumming	2014	2	0	0	3	5
Cumming	2013	2	0	0	2	4
Klugkist	2011	2	1	0	1	4
Garcia-Perez	2017	0	0	8	0	8
van Helden	2016	0	0	1	0	1
Perezgonzalez	2014	2	0	0	1	3
Citrome	2011	1	0	0	1	2
Konijn	2015	2	3	0	1	6
Hasley	2015	4	0	0	1	5
Chang	2017	1	0	3	0	4
LeBel	2011	5	0	0	4	9
Miller	2017	1	0	4	1	6
Lambdin	2012	5	14	0	1	20
Hofmann	2011	2	0	0	0	2
Perezgonzalez	2015	0	3	0	2	5
Szucs	2017	7	12	0	11	30
Savalei	2015	0	0	3	3	6
Bradley	2016	3	1	0	1	5
Baird	2016	4	1	0	4	9
Garamszegi	2017	7	2	0	1	10
Martin	2014	1	1	0	1	3
Hubbard	2011	1	1	0	0	2
Johanson	2011	6	0	0	1	7
Engman	2013	3	5	0	0	8

Note. All extracted arguments, solutions, and conclusions are claims made by the authors of the papers and are not necessarily valid.

The perspective on NHST from this zone finds a clear expression in Szucs and Ioannidis (2017, p.13):

> NHST [...] does not allow for systematic knowledge accumulation. In addition, both because of its shortcomings and because it is subject to major misunderstandings it facilitates the production of non-replicable false positive reports. Such reports ultimately erode scientific credibility and result in wasting perhaps most of the research funding in some areas.

In summary of this zone, the deficiencies and their resulting misconception concerning NHST are severe to the extent that scientific progress is impeded to dangerous levels and radical methodological change is required. The online available data (https://osf.io/bktfx/) gives a complete breakdown of the three zones from their analytic themes and subthemes to descriptive themes (i.e., arguments, solutions, conclusions).

Discussion

Summary

As this systematic review shows, NHST and its practice are diverse and intricate. We extracted data from 42 papers published between 2011 and 2018 in psychology and psychological methods journals. Combined, the papers discussed 39 deficiencies of NHST, 31 misconceptions or misuses, 20 arguments in defense of NHST, and 33 solutions to NHST problems in particular or statistical inference problems in general. Across the papers, these 123 items were supported by 508 unique pieces of evidence (i.e., references to other papers, examples, simulation studies, or mathematical proofs). The 123 arguments and solutions are employed to reach 14 types of conclusions. Collectively, these 137 items make up the landscape NHST within which the author's positions, or perspective if you will, on NHST – in terms of its qualities, problems, and their solutions – can be interpreted as the area of the landscape that is visible to the author from their point of view.

Within this landscape, we identified general zones. Based on their content, the arguments, solutions, and conclusions can be synthesized into three overarching zones enclosing certain practices and problems of NHST. The mild zone houses that which is actually in favor of NHST and only encloses problems in and solutions for policy, education, human psychology, and research culture. The moderate zone contains acknowledgments of some defects and misuse of NHST in combination with structural but not fundamental change in statistical inference (i.e., continue with frequentist hypothesis testing in a different form than NHST). The hazardous zone incorporates fundamental deficiencies and a systematic desire in the researchers that NHST will never be able to satisfy, in any shape or form. In this case, a complete break from NHST is required and a different statistical inference method needs to be adopted.

Embedding and Implications

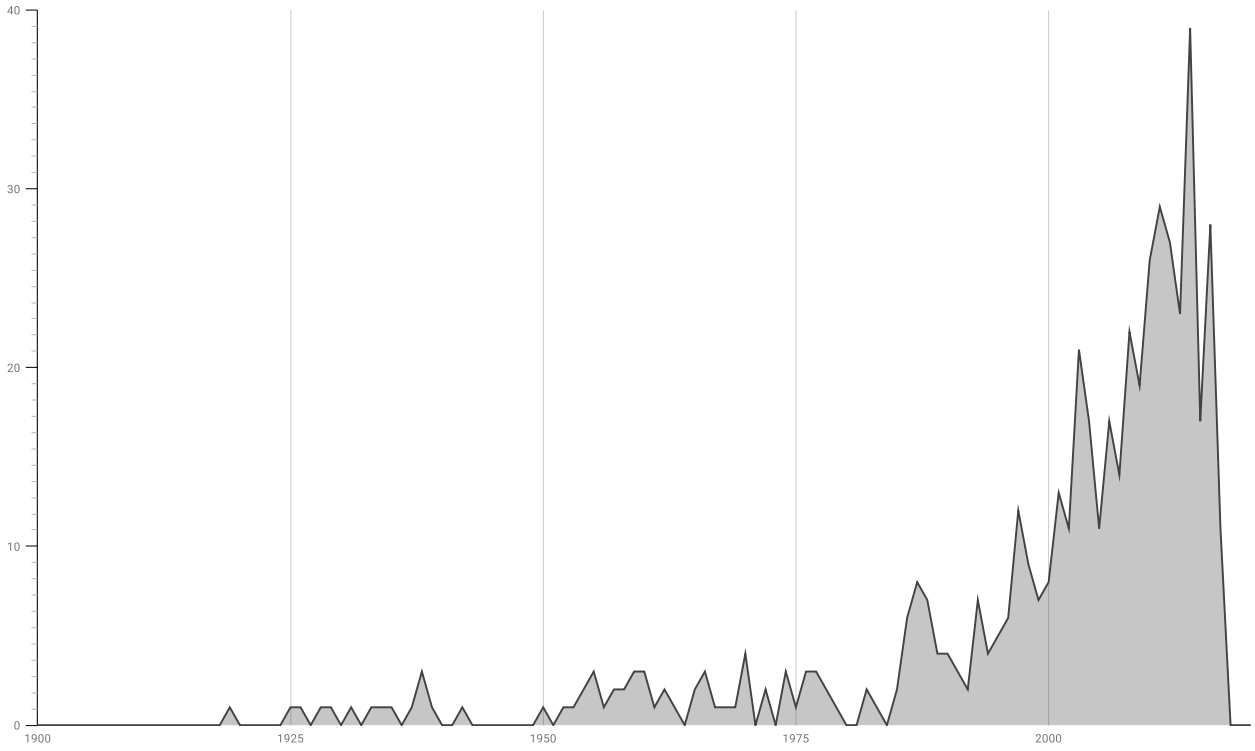
At the moment, the replication crisis is not (yet) considered solved, there is still fierce debate on how to address the misuse of NHST, and disagreement continues on if and how NHST should be changed or replaced by which alternative. The contribution of our systematic review to this situation is twofold.

Firstly, our results offer a possible explanation of why these debates have not been resolved. The NHST landscape is vast. Vast enough that no single paper on the topic provides a complete overview of NHST's machination, its application, its consequences, and how it is perceived. There is more than enough space for its inhabitants to hold incompatible views without the threat of devastating conflict, because their perspectives do not or minimally overlap and they have the internal coherence of their standpoint to fall back on. For instance, the proponents of open science (Freese & King, 2018; Lindsay et al., 2016; Munafò, 2016; van 't Veer & Giner-Sorolla, 2016) build their position on arguments concerning human psychology and research culture (i.e., how research is being done), which entail the improvement of replication rates. From another position in the landscape, these solutions are insufficient, because the inference method is considered at fault and these superficial cosmetic fixes only give a false sense of betterment (e.g., Szollosi et al., 2019). However, they are too far apart for a deciding conflict to ensue where one emerges as the victor. Their positions are not built on the same or neighboring arguments; thus, their opposing conclusions do not threaten that on which it is founded and conflicts remain unresolved. Conflicts that do concern these foundations would be dialectic: each needs to improve or move (i.e., change) their positions in response to the other's attack, eventually leading to surrender/victory or consensus. Without such threats to one's position, one's position remains safe; out of the reach of the arguments of others and where there is the internal consistency of one's own arguments to fall back.

As a result, published denouncements of each other's positions do not have to be perceived as carrying the force that would motivate change in one's position. As another example, there are several many-authors papers that call for different changes in the application of statistical methods. One such paper suggests leaving

Figure 3

Distribution of publications dates of the pieces of evidence extracted from the papers. These pieces of evidence were referenced in the papers to support the particular claims that were extracted as the descriptive themes.



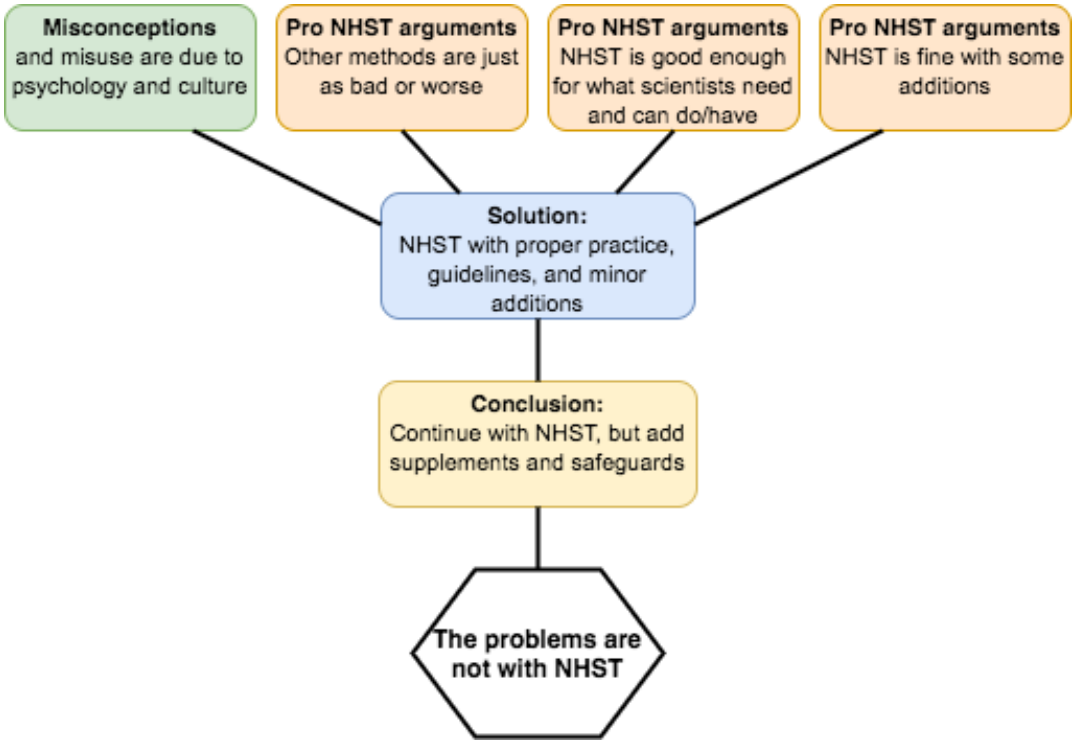
NHST as it is, but only lowering the threshold of statistical significance (Benjamin et al., 2018)[6], which is consistent with our mild zone. Another suggests taking both Type I and Type II error rates into account and have scientists set them with respect to the particularities of their research aim and subject (i.e., the Neyman-Pearson approach to frequentist hypothesis testing; Lakens et al., 2018), which is consistent with our moderate zone. A third paper suggests abandoning hypothesis tests and thresholds in favor of continuous measures of estimation and model adequacy (McShane et al., 2019), which is consistent with our hazardous zone. Hence, these papers draw on different parts of the NHST landscape (i.e., arguments). As a result, they argue against each other without really affecting the position of their opponents. Consequently, as long as those trying to resolve these problems do not consider each other's position in the NHST landscape and do not target those areas where their conflicting positions overlap, it is likely that conflicts will not be threatening enough to motivate defense or change. More specifically, researchers who are being critiqued on their NHST positions or want to critique others can identify their position and that of others within the landscape, hopefully giving them insight into the broader themes the critique is or should be nested in?

Secondly, our data can be used to move towards a resolution of these debates on how scientific practice should be improved. The overview our data provides could serve as a starting point for statisticians and methodologists to formulate alternatives that are more likely to have the support from a large group of researchers. Moreover, it provides a neutral classification for academic researchers using NHST to reflect upon their own position and account for the advantages and drawbacks of using NHST. More particular, our results can be used to identify where in the NHST landscape particular positions overlap. The arguments and solu-

[6]It needs to be noted that (many of) the authors actually prefer Bayesian hypothesis tests, but see this solution as more feasible for the near future. However, taking this into account does not change the argument.

Figure 4

The mild zone in the NHST landscape: NHST itself is not the problem. This zone consists of three NHST defenses themes (orange), one misconception theme (green); one solution theme (blue), and one conclusion theme (yellow). 14 of the 42 papers ascribe to a conclusion that is part of this zone. There are an additional four papers that contain at least one misconception and at least one solution that are part of this zone. There are seven papers that only have one or more misconceptions, one or more deficiencies, or one or solutions that are part of this zone while also containing descriptive themes belonging to the other phenomena of interest whose analytic theme(s) are part of another zone (these papers are: Bradley and Brand, 2016; Cumming, 2013, 2014; Engman, 2013; Gelman, 2013; Konijn et al., 2015; Trafimow and Marks, 2015.)



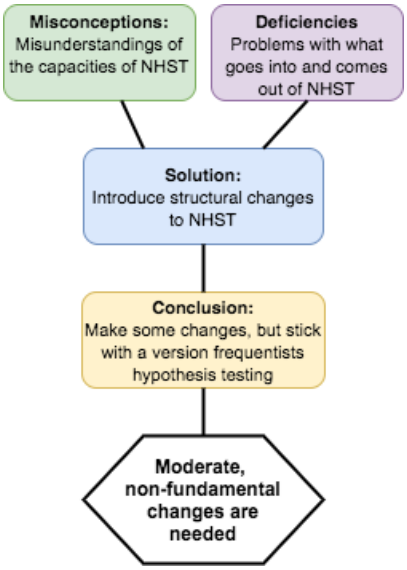
tions in these overlapping areas could be the focus of evaluation or the subject of empirical research. This would allow a systematic comparison and assessment of the particular debate positions.

In addition, our data provides some indication of relevance and interconnection. Some deficiencies and misconceptions are more prevalent in the papers than others. One explanation for this prevalence is that these problems are more relevant and severe and thus garner more attention, at least according to the authors of the included papers. Within the papers and our zones, these potentially relevant problems are connected to particular solutions, which could therefore be favored over their alternatives. In particular and as an example, educators in (more senior) methodology courses could use our data (e.g., Table A in Appendix A) as reference material to introduce the NHST debate or as embedding to discuss the various perspectives that scientists have on statistical inference. In more general terms, our material could be of assistance to those intent on improving the social sciences (e.g., policy-makers, journal editors, educators, methodologists, etc.) when deciding on how to allocate limited resources. For instance, those working towards the improvement of scientific practice through better statistics education could benefit from selecting those misconceptions of NHST that are most prevalent and evaluate if these curricular changes reduce these misconceptions.

Furthermore, an overview such as this could prove useful for those developing research methods and want to take its adequate application by fallible human scientists into account. Even though scientists might be perceived by lay people and other scientists as more rational, objective, and open-minded than non-scientists (Veldkamp et al., 2017), they are not infallible. As our data shows, there are many misconceptions and misuses possible when it comes to statistical inference via NHST, which might result from psychological biases and

Figure 5

The moderate zone: Moderate, non-fundamental changes are needed. This zone consists of one deficiency theme (purple), one misconception theme (green); one solution theme (blue), and one conclusion theme (yellow). 9 of the 42 papers ascribe to a conclusion that is part of this zone. There are an additional six papers that contain at least one misconception, deficiency or solution that is part of this zone and also one item that is part of another theme in this zone. There are ten papers that only have one or more misconceptions, one or more deficiencies, or one or more solutions that are part of this zone while also containing descriptive themes belonging to the other phenomena of interest whose analytic theme(s) are part of another zone (these papers are: Chen et al., 2017; Garamszegi and de Villemereuil, 2017; Gelman, 2013; Häggström, 2017; Halsey et al., 2015; Hupé, 2015; Johansson, 2011; LeBel and Peters, 2011; Perezgonzalez, 2014; Trafimow, 2013.



research cultures that do not incentivize methodological rigor. Also, research shows that reporting errors in statistical results are prevalent in published papers (e.g., Veldkamp et al., 2014). It is thus advisable that these shortcomings of scientists should be taken into account, to which our overview can be of assistance, when developing or implementing alternatives to NHST with the aim of improving the quality of scientific inference.

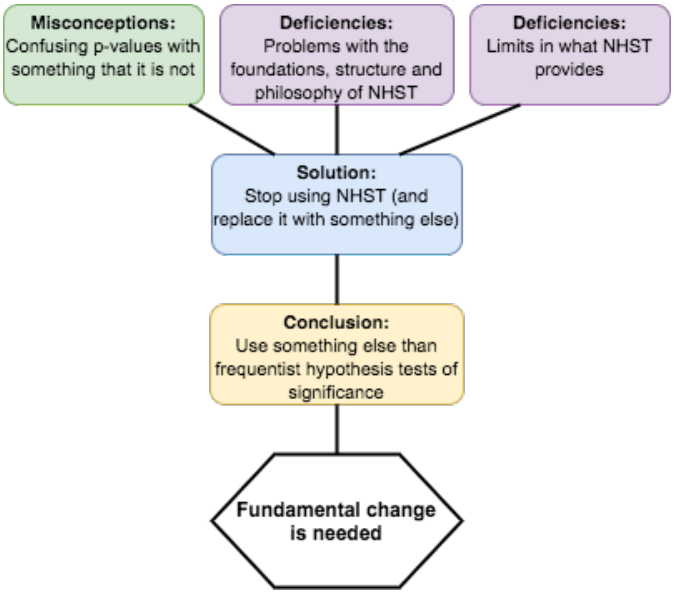
Limitations

This review suffers from several limitations. First and foremost, the methodology for the systematic review of opinion literature is not yet fully developed. We adopted methods for mixed-methods and qualitative systematic reviews, which offered guidelines. However, on several occasions, choices had to be made that had no precedent in the current methodological literature. The most prominent of these choices were the exclusion of empirical papers; using claims in the papers as our data; and categorizing the phenomena of interest in deficiencies, misconceptions, NHST defenses, solutions, and conclusions. Empirical studies of NHST and its use might also contain arguments and solutions; other types of text fragments could have been used as data; and the phenomena of interest could have been combined, further subdivided, or differently delineated. For instance, the requirement of a positive element or alternative in a proposed solution would have precluded 'stop using NHST' from being an instance of that phenomenon of interest. Different choices would have led to different results. However, without proper methodological guidelines, it is impossible to predict if these alternative choices would have led to meaningful differences in results and if these alternative results would be better or worse.

Secondly, a quality assessment of the included literature is lacking. It is general practice for systematic reviews to evaluate the quality of the data and either only include data that meet certain quality criteria (e.g., Tong et al., 2012) or conduct sensitivity analysis of the review results with respect to the quality assessments of the data (e.g., Patsopoulos et al., 2008). However, at the moment, there is no tool available for the evaluation of the quality of particular arguments, nor the structure of a set of arguments to their conclusion. Thus, we included all arguments even though some would con-

Figure 6

The hazardous zone on NHST: fundamental change is needed. This zone consists of two deficiency themes (purple), one misconceptions theme (green); one solution theme (blue), and one conclusion theme (yellow). 19 of the 42 papers ascribe to a conclusion that is part of this zone. There are an additional five papers that contain at least one misconception, deficiency, or one solution that is part of this zone and also on item that is part of another theme in this zone. There are ten papers that only have one or more misconceptions, one or more deficiencies, or one or solutions that are part of this zone while containing any descriptive themes belonging to the other phenomena of interest whose analytic theme(s) are part of another zone (these papers are: Baird & Harlow, 2016; Chang, 2017; Garamszegi & de Villemereuil, 2017; Häggström, 2017; Haig, 2017; Lu & Belitskaya-Levy, 2015; Martin & Liu, 2014; Perezgonzalez, 2014, 2015; Rouse, 2016).



sider them to be clearly false or of such a low quality that they should not be included in the overview. When comparing the arguments, some clearly contradict each other, such as Miller claiming that "the heart of NHST is a simple, intuitive, and familiar "common sense" logic" (Miller, 2017, p.664) while others claim that NHST is incoherent (and thus illogical). This is a clear indication that at least one of them must be wrong. In such a case, a quality assessment might be able to arbitrate between conflicting cases.

Thirdly, although the identified zones within the NHST landscape generally overlap with the perspectives of the authors, there is a lack of complete correspondence. However, it does not have to be unexpected that some authors use arguments or solutions across more than one zone. It is not necessary that the authors' positions are exclusively dependent on what is present in the landscape. For instance, personal agenda, position within their academic field, or their particular education might play a role as well. To extend the landscape metaphor: borders are not always drawn along natural structures, such as rivers or mountains. Logistic, political, historical, and other factors also play a role in the position of borders between countries. The same might be said of what is included or excluded in the area occupied by an author in the NHST landscape. Consequently, this does not necessarily have to be considered as a limitation factor, but could be seen as highlighting the diversity and intricacies of the NHST landscape.

Finally, the scope of our review is limited. Time and resources required constraints. We only included essay and opinion literature from psychology and methodological journals, published between 2011 and 2018. The replication crisis and the consequent attention for methodological evaluation and reform made this, in our opinion, the most promising boundaries to work in. However, other fields might have a different perspective on NHST. Specifically, researchers in those fields might differ from psychologists in their misconceptions and misuses. Or, the authors of NHST papers in these fields might emphasize deficiencies of NHST that are ignored by authors of such papers working in psychology and vice versa. Furthermore, different deficiencies and misconceptions might have been in the spotlight before the

advent of the replication crisis in 2011 and it could be possible that different problems are addressed in empirical literature concerning NHST. In short, there is ample room for broadening and deepening the understanding of NHST in particular and statistical inference in general via systematic literature review.

Suggestions for further Research

Our paper offers handholds for the extension of our review and the development of systematic review methodology for essay and opinion literature. The access we provide to our methods and progress notes will not only allow others to evaluate our work, but also extend and improve it. Specifically, this review can be extended to other academic fields and further back in time. These extensions will allow us to gain insight into the differences and similarities between fields and across time on the perspectives and practice of NHST. Such a nuanced overview will be of value in the continuing debate on how statistical inference is and should be practiced.

This paper also counts as a proof of principle. As far as we know, this is the first systematic review of essay and opinion literature. Its successful completion, apart from the quality control, is an indication that such reviews are possible in principle. We expect that our methods can be developed into a systematic review methodology with general application, which itself can be investigated and evaluated. The academic literature is teeming with essays on a wide variety of topics. Many of them have a shared topic and argue against or support each other. At the moment, essays concerning open science, the replication crisis, and preregistration and methodological rigor are prime examples of topics on which many such essays are published. When further developed, an adequate systematic review methodology could be used to catalog the copious amounts of such papers and get a clear perspective on topics that are of clear current relevance.

Concluding remarks

In this paper, we provided an overview on the perspectives on NHST and its practice. The extensive dataset that resulted from our review contains more information than we could properly describe and analyze in one paper. Thus we conclude with the wish that others might use our data and further investigate this fascinating topic.

Author Contact

Corresponding author: Noah van Dongen, nnnvandongen@gmail.com

ORCID - Noah van Dongen: 0000-0003-0387-7388
ORCID - Leonie van Grootel: 0000-0001-6675-9018

Conflict of Interest and Funding

There are no conflicts of interest to declare. The research by NvD was supported by Starting Investigator Grant No. 640638 (“OBJECTIVITY – Making Scientific Inferences More Objective”) of the European Research Council (ERC).

Author Contributions

Noah van Dongen was responsible for the conceptualization, data collection, analysis, and writing the original draft of the article. Leonie van Grootel was responsible for supervision, assisted with the conceptualization and analysis, and reviewed and edited the article.

Open Science Practices

This article earned the Open Data and Open Materials badge for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Baird, G. L., & Duerr, S. R. (2016). Reflections concerning recent ban on NHST and confidence intervals. *Journal of Modern Applied Statistical Methods*, *15*(2), 821–824.

Baird, G. L., & Harlow, L. L. (2016). Does one size fit all? A case for context-driven null hypothesis statistical testing. *Journal of Modern Applied Statistical Methods*, *15*(1), 100.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454.

Baril, G. L., & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist*, *50*, 1098–1099.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., & Berk, e. a., Richard. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*(219), 325–335.

Booth, A. (2006). "Brimful of STARLITE": Toward standards for reporting literature searches. *Journal of the Medical Library Association*, *94*(4), 421–429.

Bradley, M. T., & Brand, A. (2016). Significance testing needs a taxonomy: Or how the Fisher, Neyman–Pearson controversy resulted in the inferential tail wagging the measurement dog. *Psychological Reports*, *119*(2), 487–504.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., & Heikensten, e. a., Emma. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Campitelli, G., Macbeth, G., Ospina, R., & Marmolejo-Ramos, F. (2017). Three Strategies for the Critical Use of Statistical Methods in Psychological Research. *Educational and Psychological Measurement*, *77*(5), 881–895. https://doi.org/10.1177/0013164416668234

Carroll, C., Booth, A., & Lloyd-Jones, M. (2012). Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. *Qualitative Health Research*, *22*(10), 1425–1434.

Center for Open Science. (2020). Registered reports. Retrieved September 22, 2020, from https://www.cos.io/initiatives/registered-reports

Chang, M. (2017). What constitutes science and scientific evidence: Roles of null hypothesis testing. *Educational and Psychological Measurement*, *77*(3), 475–488.

Chen, G., Taylor, P. A., & Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *NeuroImage*, *147*, 952–959.

Citrome, L. (2011). The Tyranny of the P-value: Effect Size Matters. *Klinik Psikofarmakoloji Bülteni-Bulletin of Clinical Psychopharmacology*, *21*(2), 91–92. https://doi.org/10.5455/bcp.20110706020600

Cohen, J. (1994). The world is round (p<.05). *American Psychologist*, *49*(12), 997–1003.

Cohen, J. (1995). The Earth Is Round (p < .05): Rejoinder. *American Psychologist*, *50*(12), 1103.

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qualitative health research*, *22*(10), 1435–1443.

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on psychological science*, *3*(4), 286–300.

Cumming, G. (2013). The new statistics: A how-to guide. *Australian Psychologist*, *48*(3), 161–170.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*(1), 207–218.

Engman, A. (2013). Is there life after P<0.05? Statistical significance and quantitative sociology. *Quality & Quantity: International Journal of Methodology*, *47*(1), 257–270.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Freese, J., & King, M. M. (2018). Institutionalizing transparency. *Socius: Sociological Research for a Dynamic World*, *4*, 1–7.

Frick, R. W. (1995). A problem with confidence intervals. *American Psychologist*, *50*, 1102–1103.

Garamszegi, L. Z., & de Villemereuil, P. (2017). Perturbations on the uniform distribution of p-values can lead to misleading inferences from null-hypothesis testing. *Trends in neuroscience and education*, *8*, 18–27.

García-Pérez, M. A. (2017). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*, *77*(4), 631–662.

Gelman, A. (2013). Interrogating p-values. *Journal of Mathematical Psychology*, *57*(5), 188–189.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606.

Goddard, S., & Johnson, V. E. (2015). The Lack of Reproducibility in Research. *METODE Science Studies Journal*, *5*, 175–179. https://doi.org/17.7203/metode.83.3913

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20.

Häggström, O. (2017). The need for nuance in the null hypothesis significance testing debate. *Educational and psychological measurement*, *77*(4), 616–630.

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement*, *77*(3), 489–506.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods*, *12*(3), 179–185.

Hofmann, S. G. (2011). Some more fundamental problems in clinical research: Comment on 'Statistical significance testing and clinical trials'. *Psychotherapy*, *48*(3), 223–224. https://doi.org/10.1037/a0023198

Hogben, L. (1956). The present crisis in statistical theory. *The Incorporated Statistician*, *7*(1), 3–21.

Hubbard, R. (2011). The widespread misinterpretation of p-values as error probabilities. *Journal of Applied Statistics*, *38*(11), 2617–2626.

Hupé, J.-M. (2015). Statistical inferences under the Null hypothesis: common mistakes and pitfalls in neuroimaging studies. *Frontiers in neuroscience*, *9*, 18.

Johansson, T. (2011). Hail the impossible: P-values, evidence, and likelihood. *Scandinavian Journal of Psychology*, *52*(2), 113–125.

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do We Know What We Test and Do We Test What We Want to Know? *International Journal of Behavioral Development*, *35*(6), 550–560. https://doi.org/10.1177/0165025411425873

Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures*, *9*(4), 280–302.

Laber, E. B., & Shedden, K. (2017). Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-Values for Statisticians. *Journal of the American Statistical Association*, *112*(519), 902–904. https://doi.org/10.1080/01621459.2017.1311265

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., & Buchanan, e. a., Emily M. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171.

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

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*(4), 371–379.

Lehrer, J. (2010). Feeling The Future: Is Precognition Possible? [Accessed on 11 August 2020]. https://www.wired.com/2010/11/feeling-the-future-is-precognition-possible/

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. *APS Observer*, *29*(10).

Lu, Y., & Belitskaya-Levy, I. (2015). The debate about p-values. *Shanghai Archives of Psychiatry*, *27*(6), 381–385.

Ma, Z., Pan, Y., Yu, Z., Wang, J., Jia, J., & Wu, Y. (2013). A quantitative study on the effectiveness of peer review for academic journals. *Scientometrics*, *95*(1), 1–13.

Marsman, M., & Wagenmakers, E. J. (2017). Three Insights from a Bayesian Interpretation of the One-Sided P Value. *Educational and Psychological Measurement*, *77*(3), 529–539. https://doi.org/10.1177/0013164416669201

Martin, R., & Liu, C. (2014). A note on p-values interpreted as plausibilities. *Statistica Sinica*, 1703–1716.

McGraw, K. O. (1995). Determining false alarm rates in null hypothesis testing research. *American Psychologist*, *50*, 1099–1100.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports*, *71*(2), 339–467.

Miller, J. (2017). Hypothesis testing in the real world. *Educational and Psychological Measurement*, *77*(4), 663–672.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on cumming. *Psychological Science*, *25*(6), 1289–1290.

Munafò, M. (2016). Open science and research reproducibility. *ecancermedicalscience*, *10*, 10:ed56.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Parker, S. (1995). The "difference of means" may not be the "effect size". *American Psychologist*, *50*, 1101–1102.

Patsopoulos, N. A., Evangelou, E., & Ioannidis, J. P. A. (2008). Sensitivity of between-study heterogeneity in meta-analysis: Proposed metrics and empirical evaluation. *International Journal of Epidemiology*, *37*(5), 1148–1157.

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. *Theory & Psychology*, *24*(6), 852–859.

Perezgonzalez, J. D. (2015). The meaning of significance in data testing.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712.

Rao, C. R., & Lovric, M. M. (2016). Testing point null hypothesis of a normal mean and the truth: 21st century perspective. *Journal of Modern Applied Statistical Methods*, *15*, 2–21.

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

Rouse, S. V. (2016). Of teacups and t tests: Best practices in contemporary null hypothesis significance testing. *Psi Chi Journal of Psychological Research*, *21*(2), 127–133.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological bulletin*, *57*(5), 416.

Savalei, V., & Dunn, E. (2015). Is the Call to Abandon P-Values the Red Herring of the Replicability Crisis? *Frontiers in Psychology*, *6*, 1–4. https://doi.org/10.3389/fpsyg.2015.00245

Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard Psychology literature with Registered Reports [Accessed on 14 August]. https://psyarxiv.com

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*(1), 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*(11), 1359–1366.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is preregistration worthwhile. *Trends in Cognitive Sciences*, *24*(2), 94–95.

Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*, 390.

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). EPPI-Reviewer: Advanced software for systematic reviews, maps and evidence synthesis.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1), 45.

Tong, A., Flemming, K., McInnes, E., Oliver, S., & Craig, J. (2012). Enhancing transparency in reporting the synthesis of qualitative research: Entreq. *BMC Medical Research Methodology*, *12*(1), 181.

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

Trafimow, D. (2013). Descriptive vs. inferential cheating.

van Helden, J. (2016). Confidence Intervals Are No Salvation from the Alleged Fickleness of the P Value. *Nature Methods*, *13*(8), 605–606. https://doi.org/10.1038/nmeth.3933

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology — A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12.

Veldkamp, C. L. S., Hartgerink, C. H. J., van Assen, M. A. L. M., & Wicherts, J. M. (2017). Who believes in the storybook image of the scientist? *Accountability in Research*, *24*(3), 127–151.

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One*, *9*(12), e114876.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld

& I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123–138). John Wiley & Sons.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilcox, R. R., & Serang, S. (2017). Hypothesis Testing, p Values, Confidence Intervals, Measures of Effect Size, and Bayesian Methods in Light of Modern Robust Techniques. *Educational and Psychological Measurement*, *77*(4), 673–689. https://doi.org/10.1177/0013164416667983

Wilson, E. B. (1923). The statistical significance of experimental data. *Science*, *58*(1493), 93–100.