

The New Statistics for applied linguistics

Gerben Mulder

VU Amsterdam

The New Statistics is an approach to scholarly research which offers an alternative to the problematic overreliance on significance testing currently plaguing the research literature. This paper describes the problems associated with significance testing and introduces the key concepts of the data-analysis that best fits with the goals of the New Statistics: estimation of effect sizes and confidence intervals. These concepts will be applied in a reanalysis of the summary data from an article that was recently published in this journal. This makes it possible to compare the estimation approach advocated by the New Statistics to the standard significance tests and to discuss potential drawbacks of this approach as a means of gathering quantitative evidence in support of our substantive hypotheses.

Keywords: New Statistics, estimation, confidence intervals, Bayesian Credible Intervals, significance testing, p -value

1. Introduction

In early 2019 a remarkable event occurred: a comment was published in *Nature* calling for the retirement of statistical significance (Amrhein, Greenland, & McShane, 2019). Contrary to what the reader might assume, the remarkable part is not that the authors want to get rid of statistical significance – there are very good reasons for abandoning it; what is surprising is that the call is endorsed by more than 850 scholars from many different academic fields.

Just a few years earlier another noteworthy publication had appeared: the American Statistical Association (ASA) issued a consensus statement on the favorite measure of many researchers: the p -value (Wasserstein & Lazar, 2016). This was remarkable because statisticians are not known for making statements about the use of statistical measures in scholarly work. Yet this was exactly what the statement was about, although it was more about the misuse than the use of the p -value in scientific studies.

The scientists' call to abandon statistical significance and the statisticians' statement about the p -value can be seen as the consequences of a deep concern that the current, standard method of statistical analysis, i.e. rejection (or not) of a statistical null-hypothesis on the basis of the p -value, leads to a distortion of the scientific process. This concern is not new. On the contrary, there is a long history of critique of significance testing (see among many others Berkson, 1942; Bakan, 1966; Carver, 1978; Cohen, 1994; Gigerenzer, 2004; Hubbard & Lindsay, 2008; Lambdin, 2012; Ziliak & McClosky, 2008; Meehl, 1978, 1997; Roozeboom, 1960; Schmidt, 1996). In fact, as McShane, Gal, Gelman, Robert, and Tackett (2019) note, there is such a large amount of literature on the topic that it is infeasible to provide a complete overview in a paper such as this one; but see for instance Nickerson (2000) for a thorough review of the significance testing controversy.

The recent heightened interest in the topic may be related to the replication crisis in psychology (Chambers, 2018; Cumming, 2014), other social sciences, and biology (McShane et al., 2019). This replication crisis may in part be the result of an unfortunate combination of overreliance on statistical significance testing, inadequate understanding of the limitations of the procedure, and researchers frequently misinterpreting the results of statistical tests (Bakan, 1966; Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Haller & Kraus, 2002; Hubbard, 2004; Oakes, 1986; see also Mulder, 2016, 2019 for an elementary discussion).

The problems associated with interpretation of statistical significance – or lack thereof – are especially apparent in the context of a procedure called null-hypothesis significance testing (NHST) (Gigerenzer et al., 1989; Hubbard, 2004).¹

The basic form of NHST is as follows (Gigerenzer & Marewski, 2015; Kline, 2013). In order to test a substantive research hypothesis, one temporarily assumes that a statistical null-hypothesis is true. Usually this is a nil-hypothesis, e.g. that the difference between population means or the value of a population correlation is exactly equal to zero. The statistical hypothesis corresponding to the actual research hypothesis is generally not specified. A statistical test is performed in order to determine whether the statistical null-hypothesis can be rejected or not.

The p -value of the statistical test plays a central role in NHST. It functions both as a test statistic and as a measure of support. The p -value as test statistic is used to determine whether the null-hypothesis can be rejected. The criterion for rejection is usually 5%. This means that only if the p -value of the statistical test is smaller than 5% (a statistically significant result), the null-hypothesis is rejected and the research hypothesis accepted. The p -value is also used as a measure of support: The smaller the p -value ('the more significant') the more emphatically the results support the research hypothesis.

1. The term NHST is used to refer to NHST as it is commonly practiced, for instance in this journal.

Intuitively, NHST appears impressive. Indeed, the procedure seems to provide a way of deciding on the tenability of our research hypotheses in the light of our data and at the same time seems useful for assessing the extent to which the decision is justified. Sadly, it cannot do anything of the sort. The procedure is neither a test of a substantive research hypothesis, nor does it provide a measure of support for a substantive hypothesis. Rather, it is a test of a statistical null-hypothesis against an unspecified statistical alternative, and the p -value is not a measure of support for the (statistical) alternative hypothesis, but (presumably) a measure of inductive evidence against the null-hypothesis (Hubbard, 2004; Gigerenzer, et al., 1989; Perezgonzalez, 2015), albeit one that overstates that evidence (Berger & Sellke, 1987; Hubbard & Lindsay, 2008; Rouder, Speckman, Dongchu, Morey, & Iversen, 2009). From an epistemological perspective, then, using NHST to assess the support for substantive hypotheses is untenable (Hubbard & Lindsay, 2008; Meehl, 1978, 1997; Roozeboom, 1960).

Even if one does not accept the claim that NHST is epistemologically problematic, the frequently occurring misinterpretations of significance tests may also lead one to doubt the scientific value of NHST. ‘Significant’ results (i.e. $p < .05$) are commonly erroneously interpreted as meaning that the null-hypothesis is likely to be false, that the results are probably not due to chance, that the result is likely to replicate and that it is unlikely that rejection of the null-hypothesis is an error (Kline, 2013). Likewise, ‘non-significance’, (i.e. $p > .05$) is taken to mean that the null-hypothesis is likely to be true, that the results are due to chance, or even that the research hypothesis is probably false. These misunderstandings may not only lead to overconfidence in claims about the (non)existence of effects or relations, but may also lead to practical advice that is simply not justified by the statistical analyses of the relevant observations. That is to say, it may well be sound advice, but the point is that the soundness of the advice does not follow from the results of a significance test.

Given the fact that NHST is mistakenly seen as a procedure for testing substantive hypotheses in combination with the frequent misunderstandings of the results of significance tests, it should not come as a surprise that in published work we frequently find conclusions that are not supported by the statistical analysis of the data. Neither should it be surprising that many researchers and methodologists take issue with significance testing of null-hypotheses and the important role that this plays in our academic work. That is why statistics reformers, in an attempt to improve the quality of our research, do not only point out common misconceptions surrounding significance testing, but also provide alternatives to NHST (Kline, 2013). A recent issue of the *American Statistician* (Wasserstein, Schirm, & Lazar, 2019), for instance, contains more than 40 contributions discussing alternatives to significance testing.

This article will introduce one of the alternatives to NHST: the estimation approach (also called the ‘New Statistics’: Calin-Jageman & Cumming, 2019a; Cumming, 2012, 2014; Kruschke & Liddell, 2018).² The estimation approach will be introduced and illustrated by statistically (re)analyzing a recently published article in the *Dutch Journal of Applied Linguistics* (Van Hilten & Van Vuuren (2017)).³

The remainder of this contribution is as follows. First, the research reported in the recent article and the interpretation of the statistical results will be discussed. Second, the key ideas of the estimation approach as described by Calin-Jageman and Cumming (2019a) will be illustrated, by applying them to the reported data. Third, two interpretations of the 95% CI will be considered in more detail. These three steps provide an illustration of the interpretation of the estimation results of such a study from a frequentist and a Bayesian perspective, and an opportunity to contrast these perspectives with the standard (but often incorrect) conclusions following a significance test.

2. Does it feel non-native?

Van Hilten and Van Vuuren (2017) investigate the substantive research hypothesis that the use of clause-initial place adverbials in English texts written by advanced Dutch students of English as a foreign language leads native speakers of English to judge these texts as less coherent, continuous and native-like than texts written by native speakers of English.

Van Hilten and Van Vuuren (2017) provide the following examples:

- (1) a. Fitzgerald’s bekendste roman, uitgegeven in 1925, is *The Great Gatsby*.
In deze roman wordt de American Dream bekritiseerd.
- b. Fitzgerald’s most famous novel, published in 1925, is *The Great Gatsby*.
In this novel the American Dream is criticized.
- c. Fitzgerald’s most famous novel, published in 1925, is *The Great Gatsby*.
This novel criticizes the American Dream.

2. The following titles are recommended for readers with basic training in statistics who want to learn more about elementary alternatives: Cumming and Calin-Jageman (2017) and Kline (2013). Readers with more advanced statistical training might also consider Kruschke (2015), an introduction to the Bayesian alternative to the estimation approach.

3. The selection of the article is a happy coincidence. It was selected for a paper illustrating sample size planning for contrast analysis in common experimental designs, but its statistical contents happened to be perfect for illustration of the points made in the current contribution as well.

Van Hilten and Van Vuuren (2017) explain that

“Dutch learners are likely to translate (1a) with an initial place adverbial which serves as a local anchor, like [sic] in (1b). In English, on the other hand, the subject tends to function as a neutral discourse link, as in (1c). Apart from establishing a more neutral discourse link, (1c) adheres to the English principles of information structuring by placing the new information in the sentence in post-verbal end-focus position. Pragmatically infelicitous patterns like in (1b) can result in incoherence and a disruption of continuity in L2 texts (Baker, 1992, pp.120–133).” (pp.198–199)

In short, the use of the sentence-initial place adverbial is more common in texts written by a non-native speaker of English, so the presence of these adverbials may be a sign of non-nativeness, and its use may also result in incoherence and discontinuity.

The hypothesis was investigated by means of an experimental study. The authors used a repeated measures design in which participants read 4 texts: 3 non-native texts, containing 3–4 sentence-initial place adverbials, and 1 native text, containing none. In order to assess the influence of the presence of the place adverbials on participants’ judgments of nativeness, coherence and continuity, the participants responded to three 5-point semantic differentials, one for each of the three dependent variables.

Let us focus on the research question, namely whether the use of clause-initial place adverbials is perceived as non-native by native speakers of English. We will only consider the quantitative results. The authors answer this question by means of contrast analysis. The analysis proceeds by first calculating a contrast score for each participant, which is simply the difference between the native-like rating the participant gave to the control (native-writer) text and the mean of the ratings the participant assigned to the non-native texts. Second, the resulting contrast scores are used to test the statistical null-hypothesis that the population mean of the contrast scores is exactly zero. Whether this null-hypothesis can be rejected is usually tested with a t-test or F-test (Maxwell, Delaney, & Kelley, 2017; Rosenthal, Rosnow, & Rubin, 2000).

The p -value of the test reported by the authors is $p = .41$. This is considerably larger than .05, so the null-hypothesis that the population mean is exactly zero cannot be rejected. These results are interpreted as lack of evidence for the substantive hypothesis. The authors conclude:

The data provide no evidence that clause-initial place adverbials negatively affect native speakers’ perception of Dutch EFL writing with regard to nativeness, continuity, and coherence. The comparisons between the native and non-native texts revealed no significant differences. Taken together, the non-native-speaker texts

were not perceived as significantly less native-like, coherent, and continuous than the native-speaker text. (p. 208)

So, here we see an example of a statistically non-significant test result being interpreted as “no evidence in support of the research hypothesis.”⁴ Thus, the fact that the significance test does not allow the decision to reject the statistical null-hypothesis, is seen as lack of evidence in support of the substantive research hypothesis. Indeed, the authors explicitly claim that there is no such evidence.

However, this conclusion does not seem to be quite right, considering the summary statistics reported in the article. The sample means of the non-nativeness ratings for the three non-native texts are 2.47, 2.37, and 1.73, the mean value of which is 2.19, and the sample mean of the control text equals 2.00. Note that the mean nativeness rating for the three non-native texts is higher (i.e. more foreign) than the mean nativeness rating of the native control text, as the authors expected. The value of the sample mean contrast score equals $2.19 - 2.00 = 0.19$.

Now, a difference of 0.19 on the scale used by the authors is difficult to interpret, but the standardized value of the difference yields $d = 0.15$,⁵ which according to rules-of-thumb developed for psychology is considered to be a small effect (Cohen, 1988; Field, 2015).⁶ By comparison, in some subfields of communication (i.e. persuasion research) a standardized effect of around 0.15 can be considered a medium effect (Mulder, 2019).

In any case, the authors found a non-negligible effect that is perfectly in line with their substantive hypothesis, so the claim that no evidence was found does not seem to be right. It should be noted, however, that the authors’ claim is not nearly as problematic as the frequently occurring conclusions “there is no effect”, or “the research hypothesis should be rejected”, which seem to be particularly dubious when the results agree with the research hypothesis.

What these conclusions seem to have in common is that the results of the significance test are incorrectly interpreted as the amount of evidence in support

4. It may be worthwhile to point out to the reader that this interpretation of results is indicative of the many fields that use statistical significance: the authors’ interpretation is perfectly in line with standard norms in many fields and, apparently also the norms of the reviewers of the article.

5. This is an estimate of the ratio of the population value of the mean contrast score to the population standard deviation of the contrast scores. This is an analogue of Cohen’s d for dependent samples (often denoted d_z). The estimate was obtained by taking the square root of the value of F the authors provide (i.e. $F(1, 19) = 0.71$) divided by the sample size ($N = 30$): $\delta = \sqrt{(F/N)} = \sqrt{(0.71/30)} = 0.15$.

6. The population value of the standardized contrast score is $\delta = \psi/\sigma$, the expected contrast score divided by the standard deviation of the (population) contrast scores.

of a substantive research hypothesis: if we (can) decide to reject the statistical null-hypothesis ($p < .05$) there is evidence, otherwise, there is none. It can be argued that this faulty interpretation is based on failure to recognize the difference between substantive and statistical hypotheses and the different processes involved in accepting or rejecting them.

A statistical hypothesis specifies the theoretical probability distribution of the possible observations (Hacking, 1965; Polya, 1954). The statistical null-hypothesis tested by the authors, for example, is that the possible contrast scores are normally distributed with mean equal to zero and unknown standard deviation. A rejection of a statistical null-hypothesis is either a good decision or an error. If it is not an error, rejection means that an alternative statistical hypothesis (e.g. that the population mean is a value other than zero) better describes the theoretical probability distribution. But since the decision to reject may be an error, rejection of the null-hypothesis does not imply that the statistical alternative is true.

But even if the statistical alternative hypothesis were true, this would not entail that the substantive hypothesis behind it is true (Kline, 2013). Acceptance of the alternative hypothesis may provide grounds for claiming that a systematic explanation of the results is called for, but it does not imply that the best explanation is the one we hypothesized (e.g. the presence or absence of sentence initial place adverbials). For the same reasons, non-rejection of the statistical null-hypothesis does not mean that the truth of the statistical null-hypothesis has been demonstrated nor that the substantive hypothesis behind the null-hypothesis is true (i.e. that there is no effect or no systematic relation between the variables; that sentence-initial place adverbials do not make texts less native-like, for example).

Even though rejection or non-rejection of the statistical hypothesis may be justified, it seems impossible to directly translate such a result to the tenability of a substantive hypothesis. Acceptance (or rejection) of a substantive hypothesis requires evaluation of the evidence and comparison with competing substantive explanations (Kline, 2013; Roozeboom, 1960). These latter processes take place on a much higher level of abstraction than the simple yes/no-decisions involved in statistical null-hypothesis testing.

So, it is a mistake to conclude on the basis of $p > .05$ that there is no evidence in support of the research hypothesis, that there is no effect or that the research hypothesis should be rejected, because the evidence in support of the substantive hypothesis has not been evaluated. Rather, one has made the decision not to reject the statistical null-hypothesis.⁷

7. Contrary to what happens in practice, with NHST one cannot accept the null-hypothesis. The primary reason is that in general we do not know what the probability is that accepting

3. The New Statistics

The major difference between the New Statistics approach to data-analysis and NHST is that the focus is no longer on the rejection of nil-hypotheses but on estimating effect sizes, the uncertainty of those estimates, and meta-analysis (Calin-Jageman & Cumming, 2019a, 2019b; Cumming, 2012, 2014; Cumming & Calin-Jageman, 2017; Kruschke & Liddell, 2018). Or, as the title of Calin-Jageman and Cumming (2019a) succinctly states: “Ask how much, how uncertain, and what else is known.” Here we will restrict attention to the data-analysis in a single study and focus on the “how much” and “how uncertain” questions.

Data analysis from the New Statistics perspective amounts to using techniques for estimation. These techniques themselves are not new at all, but using them as the main approach to statistical inference would be new for many researchers (Cumming, 2014). The basic estimation approach can be summarized in the following five steps (these five steps are paraphrased versions of five of the seven steps described in Cumming and Calin-Jageman, 2017, p.12).

1. Express the research question as a “how much” or a “to what extent” question.
2. Choose the most appropriate measure for answering that question.
3. Develop a study design in which that measure is used to obtain point and interval estimates that provide an answer to your research question.
4. After the results are in, calculate point and interval estimates and create a figure.
5. Interpret the results of the analysis using knowledge about the research context.

Expressing research questions as “how much” or “to what extent” makes it obvious to the researcher that the study should provide high quality quantitative information. Focusing on obtaining high quality quantitative information right from the start of the study influences all decisions to be made in gathering observations and analyzing and reporting the results.

Answers to quantitative questions consist of estimates of population effect sizes and estimates of the degree of uncertainty associated with those estimates.

The uncertainty of an effect size estimate is expressed with an interval estimate, usually a 95% confidence interval (CI). Step 4 of the basic steps of the estimation approach therefore consists of calculating an effect size and a 95%

the nil is an error (a so called type II error). But if we neither accept nor reject, the decision theoretic approach of NHST fails: we are left in a state of indecision, so we do not know what to do. And knowing what to do is exactly why one would use a decision theoretic approach to statistical analysis.

CI and creating a figure illustrating these statistics (Cumming, 2012; Cumming & Calin-Jageman, 2017). Kline (2013) describes how to obtain those estimates for a wide range of effect sizes in many applied situations. Researchers who do not want to adhere to frequentist principles in data analysis may resort to other approaches. For instance, from a Bayesian perspective uncertainty can be expressed by Bayesian credible intervals (Kruschke, 2015; Kruschke & Liddell, 2018; Norouzian, De Miranda, & Plonsky, 2018).

Below, the (summary) data provided by Van Hilten and Van Vuuren (2017) will be analysed from the frequentist perspective as proposed by Cumming (2012, 2014) and Cumming and Calin-Jageman (2017). The main focus will be on estimating the unknown population effect size δ and its associated 95% CI. However, in describing possible interpretations of CIs we will also have to discuss the Bayesian approach, at least to some extent, because some of these interpretations are controversial, especially when considered from an evidential perspective.

3.1 An estimate of the population effect size and the confidence interval

Let us again focus on the research question: to what extent does the use of sentence-initial place adverbials influence nativeness judgments? A quantitative answer is provided by an estimate of the population mean contrast score. Figure 1 shows the (unstandardized) results (i.e. the sample means), the mean contrast score in the sample and the 95% confidence intervals. Note that the axis to the right is for interpreting the CI of the population mean. The dotted line from the mean of the control condition to the zero point on that axis is included to show that the contrast is defined relative to that mean.

The value of the mean contrast score (0.19) is our estimate of the population value. This so-called point estimate is our quantitative answer to the research question: we have estimated that the use of sentence initial place adverbials is associated with a 0.19 increase on the nativeness scale (remember that higher scores on that scale are less-native). As explained, standardizing the contrast estimate leads to $d = 0.15$. So, it appears that the adverbials have a small population effect on nativeness ratings: using them leads to texts that native speakers judge somewhat less native (on average) than texts without them.

Of course, we should not fool ourselves into believing that the estimate of the population value is equal to its “true” value. Indeed, our estimated value is uncertain: a new sample of participants will give a different estimate. From a frequentist perspective, the 95% CI expresses that uncertainty. For this particular effect size the 95% CI is $[-0.21, 0.51]$. This CI was obtained with the MBESS package for R (Kelley, 2007).

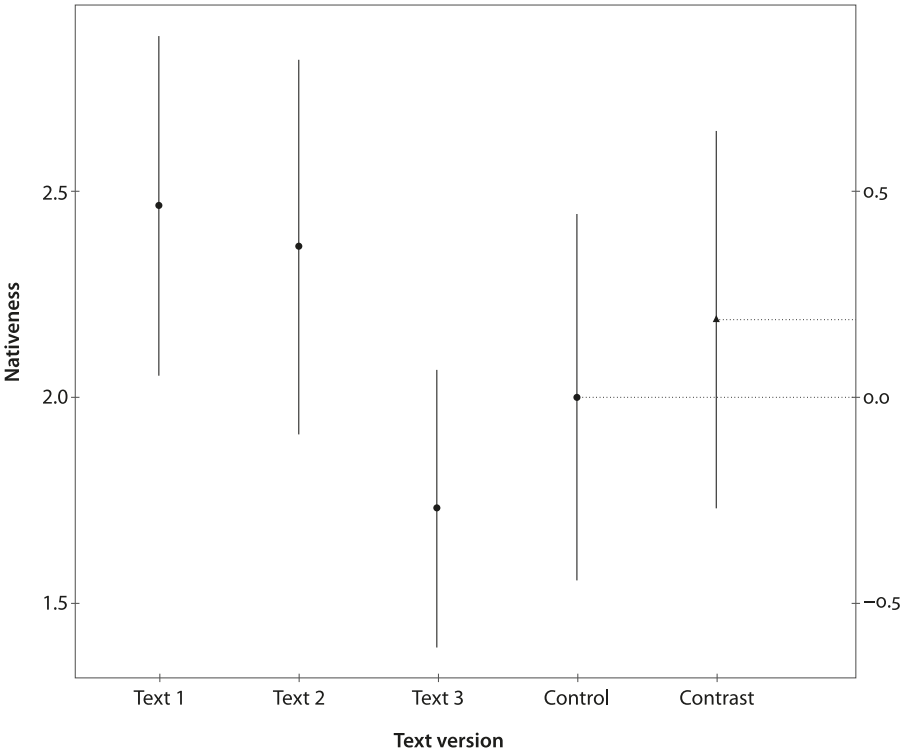


Figure 1. Estimates of population values for each text and the contrast

Following the recommendations of the APA (2010), our estimation result is summarized as follows: $d = 0.15$, 95% CI $[-0.21, 0.51]$. An alternative way of writing the interval estimate makes it more obvious why it is called an interval estimate: $-0.21 \leq \delta \leq 0.51$; we assert that the unknown value of the population effect size is between -0.21 and 0.51 . Using the rules-of-thumb for interpreting the standardized effect size, our estimate is that the effect size is between a small negative effect (texts with clause-initial place adverbials judged to be more native) and a medium positive effect (texts with clause-initial place adverbials judged to be less native). Thus, even though the point estimate indicates a population effect size in the hypothesized direction, the range of values included in the interval should temper our enthusiasm.

Reporting and evaluating the CI prevents some of the more common mistakes in NHST-practice. That is why statistics reformers propose the CI as a replacement of or an addition to the significance test. Here we can see, for instance, that considering the whole range of values in the CI makes it easy to avoid incorrectly accepting the null-hypothesis that the population effect size equals zero. We have estimated that the effect size is somewhere between -0.21

and 0.51, so there is no more reason to suppose the true value is zero than one of the many other values contained in the interval. This in turn prevents us from believing that no evidence was found, that the effect is non-existent or that our research hypothesis is incorrect.

Likewise, if the result is significant, values close to zero may still be included in the interval. In fact, if $p = .05$, the null value is at one of the limits of the interval (Cumming, 2012; Cumming & Calin-Jageman, 2017). So considering the whole range of values contained in the 95% confidence interval may prevent the misconstrual of a significant result as strong support for the substantive research hypothesis, as is common in NHST-practice.

4. Interpreting confidence intervals

Cumming (2012, 2014) presents several ways of interpreting the confidence interval obtained in a single study. We will discuss two of them. The first interpretation is that the interval is one of an infinite sequence of intervals. The second interpretation is that the interval is a range of relatively plausible values for the unknown population effect size. This plausibility gradually decreases as we move from the center of the interval to its limits and beyond. This second interpretation is methodologically controversial, as will be discussed below.

A confidence interval is a frequentist concept, so purists would say that proper interpretation of a confidence interval requires adherence to frequentist theory (Kline, 2013; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Norouzzian, De Miranda, & Plonsky, 2018). Conceptually, the frequentist interpretation is that a 95% confidence interval is a statement based on a procedure that produces correct statements 95% of the times. The statement is that the unknown population value is between the lower and upper confidence limits. Now, suppose that we construct an infinite sequence of 95% CIs, each based on a new random sample: then 95% of the CIs will be a true statement about the unknown value of the parameter. Any realized 95% CI is considered to be one of this infinite sequence of CIs.

For example, the 95% confidence interval of the population effect size in the Van Hilten and Van Vuuren (2017) study is $[-0.21, 0.51]$, which translates to the statement $-0.21 \leq \delta \leq 0.51$. Now, 95% of CIs are true statements, but whether or not this particular statement is true or false cannot be determined solely on the basis of the information provided by the interval. All we know (or rather assume) is that this particular CI is one from an infinite sequence.

Note that we already knew that before we collected the data. In that sense we learned nothing from our results. This is perfectly fine, since the procedure was not developed for learning from data: it is non-evidential (Gigerenzer et al., 1989;

Hubbard, 2004; Morey et al., 2016; Neyman, 1977). But the fact that the frequentist CI is non-evidential poses quite a challenge if we want to use the interval to evaluate the evidential support for our statistical hypothesis, which is exactly what happens in the second interpretation of the 95% CI.

According to the second interpretation, Van Hilten and Van Vuuren's (2017) result $d = 0.15$, 95% CI $[-0.21, 0.51]$ suggests that the values between -0.21 and 0.51 are relatively more plausible than values outside the interval, and moreover that the values at the limits are not as plausible as the values closer to 0.15 (Cumming, 2012).

This evidential interpretation of the CI is inconsistent with frequentist theory (Morey et al., 2016; Norouzzian, De Miranda, & Plonsky, 2018) and therefore leads to logically inconsistent methodology, which is of course unacceptable from a scientific point of view. However, from a practical point of view, the 95% CI can be seen as an approximation to Bayesian credible intervals that do allow for an evidential interpretation.

Conceptually, Bayesian 95% credible intervals give us a range of potential population values that have the highest plausibility or believability, after we have used the data to update our prior beliefs about these population values (Kruschke, 2015; Kruschke & Liddell, 2018; Norouzzian, De Miranda, & Plonsky, 2018; Wiens & Nilsson, 2017).

The software JASP (www.jasp-stats.org) can be used to calculate the 95% Credible Interval based on summary statistics (see Ly, Raj, Etz, Marsman, Gronau, & Wagenmakers, 2018). Figure 2 presents the output of the Bayesian one sample t-test for the Van Hilten and Van Vuuren (2017) data.⁸

The 95% Credible Interval is $[-0.19, 0.49]$ (see upper-right corner of the graph), suggesting that the population effect sizes between a small negative effect and a medium positive effect are most plausible, given the observed data and the prior distribution. The graph shows that values close to the center of the distribution are more credible than values at the limits or beyond the credible interval.

Note that the results are very close to the values of the 95% CI limits, and in this sense the 95% CI provides an approximation to the 95% Credible Interval. The close numerical correspondence between 95% CIs and Bayesian credible intervals occurs frequently, especially in single studies (Lindley, 2000). However, it is important that no claim of generality is being made here. Indeed, as Morey et al. (2016) note, it is unwise to assume that the difference between the numerical results of the two approaches is always small, since the results of the procedures

8. Input values $t = .8426$, $N = 30$ and the default value ($r = .707$) for the scale parameter of the Cauchy prior on the effect size (Rouder et al., 2009). See Norouzzian, De Miranda, & Plonsky (2018), for a more informative and detailed discussion of the Bayesian approach.

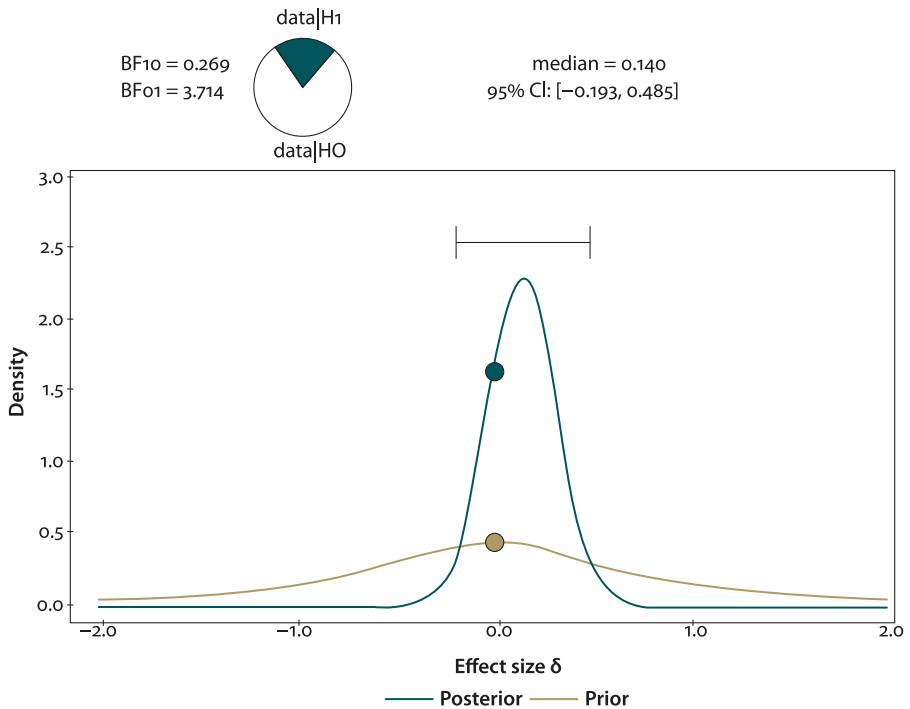


Figure 2. Output from the JASP summary statistics one sample Bayesian t-test

can lead to markedly different results, depending for instance on which procedure is used to determine the confidence limits.

The upshot of all this is that practically speaking, but certainly not philosophically, we can interpret the 95% CI as the range of most plausible population values, with plausibility decreasing as we move from the center of the interval to the limits, and beyond.

5. Conclusion

The goal of this article is to show how the New Statistics' estimation approach can be used for the analysis of applied linguistic data. Since statistics reformers promise that the estimation approach supersedes significance testing, in the sense that our scientific claims will be better justified, it is worth considering how the substantive interpretation of the estimation results differs from that of the significance test in the original article.

It was argued that the substantive claim based on the original article (i.e. that no evidence was found in support of the research hypothesis) was a mistake. It

was explained that the fact that a test of a statistical hypothesis is not significant (i.e. $p > .05$) does not mean lack of evidence in support of the substantive research hypothesis, neither does it mean that the statistical null-hypothesis or the substantive hypothesis it represents (i.e. there is no effect or relation) is true.

The estimation results suggest that sentence initial place adverbials have a small negative effect on native speakers' attitudes towards the nativeness of a text, exactly as the authors hypothesized. However, the interval estimate suggests that initial place adverbials may actually have an even more negative effect than the one obtained (the upper limit shows a medium effect), and – more worrying in the light of the substantive hypothesis – might actually increase native speakers' judgments of the (non-)nativeness of a text. So, the results are actually quite promising for the hypothesis, but the uncertainty of the estimate is so large that the results may not convince a skeptical audience. Note the striking difference between the interpretation of the significance test, i.e. “we found no evidence”, and that of the estimation results “the results are quite promising for the hypothesis”.

It is important to keep in mind, however, that strictly speaking the estimation results provide insight in plausible values of the population effect size. These candidate values are statistical hypotheses and not substantive ones. So, we still need to reason from these candidate values to substantive hypotheses and this is not a trivial task.

Let us consider a few examples. The CI suggests that the population effect size may be negative. That is, it suggests the substantive hypothesis that sentence initial place adverbials will actually increase how native-like native speakers will judge a text. Of course, this seems highly implausible, even to a non-expert. The CI also suggests that a medium effect size is plausible. But is it substantively plausible that attitudes are influenced that much? Considering the typical effect sizes in persuasive communication research, where the effects of text (message) characteristics on judgments like these play a central role, this seems somewhat far-fetched, but it is consistent with the authors' hypothesis. Could the small obtained effect size be a chance finding? Considering that zero population effect is one of the plausible candidate values, we should not immediately dismiss that possibility, but there are theoretical and empirical reasons, as the authors explain and show in their introduction, that suggest that chance may not be the most plausible explanation that comes to mind. Note that these substantive interpretations are not possible if the only information we have is that the test is not significant.

The purpose of empirical research is to assess the extent to which we should modify our beliefs in causal or other substantive hypotheses (Roozeboom, 1960), and it seems to this author that estimating effect sizes and confidence intervals (or non-frequentist alternatives) serves that purpose better than testing statistical null-hypotheses. Adopting the New Statistics leads to data analysis and results that

provide the necessary input to evaluating our substantive claims, including, of course, those based on scientific studies in the field of applied linguistics.

Acknowledgements

The author wishes to thank Susan Blackwell and two anonymous reviewers for their comments on earlier versions of this paper and their helpful suggestions for improvement.


References

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567, 305–307. <https://doi.org/10.1038/d41586-019-00857-9>
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437. <https://doi.org/10.1037/h0020412>
- Berger, O. J., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *P* Values and evidence. *Journal of the American Statistical Association*, 82, 112–122.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37, 325–335. <https://doi.org/10.1080/01621459.1942.10501760>
- Calin-Jageman, R. J., & Cumming, G. (2019a). The New Statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, 70, 271–280. <https://doi.org/10.1080/00031305.2018.1518266>
- Calin-Jageman, R. J., & Cumming, G. (2019b). Estimation for better inference in neuroscience. *ENeuro*, 6, 1–11. <https://doi.org/10.1523/ENEURO.0205-19.2019>
- Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review*, 48, 378–399. <https://doi.org/10.17763/haer.48.3.t490261645281841>
- Chambers, C. (2018). *The seven deadly sins of psychology. A manifesto for reforming the culture of scientific practice*. Princeton/Oxford: Princeton University Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). New York, NY: Academic Press.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Cumming, G. (2012). *Understanding the New Statistics. Effect sizes, confidence intervals, and meta-analysis*. New York/London: Routledge.
- Cumming, G. (2014). The New Statistics: Why and how. *Psychological Science*, 25, 7–29. <https://doi.org/10.1177/0956797613504966>
- Cumming, G. & Calin-Jageman, R. J. (2017). *Introduction to the New Statistics. Estimation, open science, & beyond*. New York/London: Routledge.
- Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5, 75–98. <https://doi.org/10.1177/0959354395051004>
- Field, A. (2015). *Discovering statistics using IBM SPSS Statistics* (4th ed.). London: Sage.

- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587–606.
<https://doi.org/10.1016/j.socsec.2004.09.033>
- Gigerenzer, G., & Marewski, J.N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41, 421–440.
<https://doi.org/10.1177/0149206314547522>
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The Empire of Chance*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511720482>
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9781316534960>
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance. A problem students share with their teachers? *Methods of Psychological Research Online*, 7, <http://www.mpr-online.de>
- Hubbard, R. (2004). Alfabet soup: Blurring the distinctions between p's and α 's in psychological research. *Theory & Psychology*, 14, 295–327.
<https://doi.org/10.1177/0959354304043638>
- Hubbard, R., & Lindsay, R.M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18, 69–88.
<https://doi.org/10.1177/0959354307086923>
- Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, 39, 979–384. <https://doi.org/10.3758/BF03192993>
- Kline, R.B. (2013). *Beyond significance testing. Statistics reform in the behavioral sciences*. Washington, DC: American Psychological Association. <https://doi.org/10.1037/14136-000>
- Kruschke, J.K. (2015). *Doing Bayesian data analysis. A tutorial with R, Jags, and Stan* (2nd ed.). London: Academic Press.
- Kruschke, J.K., & Liddell, T.M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin Review*, 25, 178–206. <https://doi.org/10.3758/s13423-016-1221-4>
- Lambdin, C. (2012). Significance tests as sorcery: Science is empirical – significance tests are not. *Theory & Psychology*, 22, 67–90. <https://doi.org/10.1177/0959354311429854>
- Lindley, D.V. (2000). The philosophy of statistics. *The Statistician*, 49, 293–337.
- Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q.F., & Wagenmakers, E.J. (2018). Bayesian reanalyses from summary statistics: A guide for academic consumers. *Advances in Methods and Practices in Psychological Science*, 1, 367–374.
<https://doi.org/10.1177/2515245918779348>
- Maxwell, S.E., Delaney, H.D., & Kelley, K. (2017). *Designing experiments and analyzing data. A model comparison perspective* (3th ed.). New York, NY: Routledge.
<https://doi.org/10.4324/9781315642956>
- McShane, B.B., Gal, D., Gelman, A., Robert, C., & Tackett, J.L. (2019). Abandon statistical significance. *The American Statistician*, 73, 235–245.
<https://doi.org/10.1080/00031305.2018.1527253>
- Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
<https://doi.org/10.1037/0022-006X.46.4.806>
- Meehl, P.E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, NJ: Erlbaum.

- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin Review*, 23, 103–123. <https://doi.org/10.3758/s13423-015-0947-8>
- Mulder, G. (2016). De kwaliteit van onderzoek. Dichotoom denken versus meta-analytisch denken. *Tijdschrift voor Taalbeheersing*, 38, 163–173. <https://doi.org/10.5117/TVT2016.2.MULD>
- Mulder, G. (2019). Een significant probleem. *Tijdschrift voor Taalbeheersing*, 41, 203–213. <https://doi.org/10.5117/TVT2019.1.014.MULD>
- Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, 36, 97–131. <https://doi.org/10.1007/BF00485695>
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301. <https://doi.org/10.1037/1082-989X.5.2.241>
- Norouzzian, R., De Miranda, M., & Plonksy, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, 68, 1032–1075. <https://doi.org/10.1111/lang.12310>
- Oakes, M. (1986). *Statistical significance*. New York, NY: Wiley.
- Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson, or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, 6, 1–11. <https://doi.org/10.3389/fpsyg.2015.00223>
- Polya, G. (1954). *Mathematics and plausible reasoning, V1–2. Induction and analogy in mathematics, patterns of plausible inference*. Princeton, NJ: Princeton University Press.
- Roozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416–428. <https://doi.org/10.1037/h0042040>
- Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research. A correlational approach*. Cambridge, UK: Cambridge University Press.
- Rouder, J. N., Speckman, P. L., Dongchu, S., Morey, R. D., & Iversen, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, 16, 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Schmidt, F. L. (1996). Statistical significance and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1, 115–129. <https://doi.org/10.1037/1082-989X.1.2.115>
- Van Hilten, M., & Van Vuuren, S. (2017). Does it ‘feel’ non-native? Native-speaker perceptions of information-structural transfer in L1 Dutch advanced EFL writing. *Dutch Journal of Applied Linguistics*, 6, 197–212. <https://doi.org/10.1075/dujal.16021.hil>
- Wasserstein, R. L., & Lazar, N. (2016). The ASA’s statement on P-values: Context, process, and purpose. *The American Statistician*, 70, 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ $p < .05$ ”. *The American Statistician*, 73, 1–19. <https://doi.org/10.1080/00031305.2019.1583913>
- Wiens, S., & Nilsson, M. E. (2017). Performing contrast analysis in factorial designs: From NHST to confidence intervals and beyond. *Educational and Psychological Measurement*, 77, 690–715. <https://doi.org/10.1177/0013164416668950>
- Ziliak, S. T., & McClosky, D. N. (2008). *The cult of statistical significance. How the standard error costs us jobs, justice, and lives*. Ann Arbor, MN: The University of Michigan Press.

Address for correspondence

Gerben Mulder
Department of Language, Literature, & Communication
Faculty of Humanities
VU Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
g.mulder@vu.nl
 <https://orcid.org/0000-0002-1569-520X>

Publication history

Date received: 18 April 2019
Date accepted: 13 March 2020
Published online: 10 September 2020