Psychon Bull Rev
DOI 10.3758/s13423-013-0572-3

BRIEF REPORT

Robust misinterpretation of confidence intervals

Rink Hoekstra · Richard D. Morey · Jeffrey N. Rouder · Eric-Jan Wagenmakers

Abstract Null hypothesis significance testing (NHST) is undoubtedly the most common inferential technique used to justify claims in the social sciences. However, even staunch defenders of NHST agree that its outcomes are often misinterpreted. Confidence intervals (CIs) have frequently been proposed as a more useful alternative to NHST, and their use is strongly encouraged in the APA Manual. Nevertheless, little is known about how researchers interpret CIs. In this study, 120 researchers and 442 students—all in the field of psychology—were asked to assess the truth value of six particular statements involving different interpretations of a CI. Although all six statements were false, both researchers and students endorsed, on average, more than three statements, indicating a gross misunderstanding of CIs. Self-declared experience with statistics was not related to researchers' performance, and, even more surprisingly, researchers hardly outperformed the students, even though the students had not received any education on statistical inference whatsoever. Our findings suggest that many researchers do not know the correct interpretation of a CI. The misunderstandings surrounding *p*-values and CIs are particularly unfortunate because they constitute the main tools by which psychologists draw conclusions from data.

R. Hoekstra (✉) · R. D. Morey · E.-J. Wagenmakers
University of Groningen, Groningen, The Netherlands
e-mail: r.hoekstra@rug.nl

J. N. Rouder
University of Missouri, Columbia, MO, USA

Keywords Confidence intervals · Significance testing · Inference

Introduction

Statistical inference is central to the justification of claims across scientific fields. When statistics such as *p*-values or confidence intervals (CIs) serve as the basis for scientific claims, it is essential that researchers interpret them appropriately; otherwise, one of the central goals of science—the justification of knowledge—is undermined. It is therefore critically important to identify and correct errors where researchers believe that a statistic justifies a particular claim when it, in fact, does not.

Of the statistical techniques used to justify claims in the social sciences, null hypothesis significance testing (NHST) is undoubtedly the most common (Harlow, Mulaik, & Steiger, 1997; Hoekstra, Finch, Kiers, & Johnson, 2006; Kline, 2004). Despite its frequent use, NHST has been criticized for many reasons, including its inability to provide the answers that researchers are interested in (e.g., Berkson, 1942; Cohen, 1994), its violation of the likelihood principle (e.g., Berger & Wolpert, 1988; Wagenmakers, 2007), its tendency to overestimate the evidence against the null hypothesis (e.g., Edwards, Lindman, & Savage, 1963; Sellke, Bayarri, & Berger, 2001), and its dichotomization of evidence (e.g., Fidler & Loftus, 2009; Rosnow & Rosenthal, 1989). In addition, it has been argued that NHST is conceptually difficult to understand and that, consequently, researchers often misinterpret test results (Schmidt, 1996). Although some researchers have

Published online: 14 January 2014

defended its usability (e.g., Abelson, 1997; Chow, 1998; Cortina & Dunlap, 1997; Winch & Campbell, 1969), there seems to be widespread agreement that the results from NHST are often misinterpreted.

For example, in a well-known study on the misinterpretation of results from NHST, Oakes (1986) presented a brief scenario with a significant *p*-value to 70 academic psychologists and asked them to endorse as true or false six statements that provided differing interpretations of the significant *p*-value. All six statements were false; nonetheless, participants endorsed, on average, 2.5 statements, indicating that the psychologists had little understanding of the technique's correct interpretation. Even when the correct interpretation was added to the set of statements, the average number of incorrectly endorsed statements was about 2.0, whereas the correct interpretation was endorsed in about 11 % of the cases. Falk and Greenbaum (1995) found similar results in a replication of Oakes's study, and Haller and Krauss (2002) showed that even professors and lecturers teaching statistics often endorse false statements about the results from NHST. Lecoutre, Poitevineau, and Lecoutre (2003) found the same for statisticians working for pharmaceutical companies, and Wulff and colleagues reported misunderstandings in doctors and dentists (Scheutz, Andersen, & Wulff, 1988; Wulff, Andersen, Brandenhoff, & Guttler, 1987). Hoekstra et al. (2006) showed that in more than half of a sample of published articles, a nonsignificant outcome was erroneously interpreted as proof for the absence of an effect, and in about 20 % of the articles, a significant finding was considered absolute proof of the existence of an effect. In sum, *p*-values are often misinterpreted, even by researchers who use them on a regular basis.

The philosophical underpinning of NHST offers a hint as to why its results are so easily misinterpreted. Specifically, NHST follows the logic of so-called frequentist statistics. Within the framework of frequentist statistics, conclusions are based on a procedure's average performance for a hypothetical infinite repetition of experiments (i.e., the sample space). Importantly, frequentist statistics does not allow one to assign probabilities to parameters or hypotheses (e.g., O'Hagan, 2004); this can be done only in the framework of Bayesian statistical techniques, which are philosophically incompatible with frequentist statistics. It has been suggested that the common misinterpretations of NHST arise in part because its results are erroneously given a Bayesian interpretation, such as when the *p*-value is misinterpreted as the probability that the null hypothesis is true (e.g., Cohen, 1994; Dienes, 2011; Falk & Greenbaum, 1995).

Within the frequentist framework, a popular alternative to NHST is inference by CIs (e.g., Cumming & Finch, 2001; Fidler & Loftus, 2009; Schmidt, 1996; Schmidt & Hunter, 1997). CIs are often claimed to be a better and more useful alternative to NHST. Schmidt, for example, considered replacing NHST by point estimates with CIs "essential for the future progress of cumulative knowledge in psychological research" (p. 115), and Cumming and Fidler (2009) argued that "NHST . . . has hobbled the theoretical thinking of psychologists for half a century" (p. 15) and that CIs address the problems with NHST, albeit to varying degrees. Fidler and Loftus stated that NHST dichotomizes researchers' conclusions, and they expected that since CIs would make precision immediately salient, they would help to alleviate this dichotomous thinking. Cumming and Finch mentioned four reasons why CIs should be used: First, they give accessible and comprehensible point and interval information and, thus, support substantive understanding and interpretation; second, CIs provide information on *any* hypothesis, whereas NHST is informative only about the null; third, CIs are better designed to support meta-analysis; and finally, CIs give direct insight into the precision of the procedure and can therefore be used as an alternative to power calculations.

The criticism of NHST and the advocacy for CIs have had some effects on practice. At the end of the previous century, the American Psychological Association (APA) installed the Task Force on Statistical Inference (TFSI) to study the controversy over NHST and to make a statement about a possible ban on NHST. The TFSI published its findings in 1999 in the *American Psychologist* (Wilkinson & TFSI, 1999), and it encouraged, among other things, the use of CIs, because "it is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual *p*-value or, better still, a confidence interval" (Wilkinson & TFSI, 1999, p. 599). The advice of the TFSI was partly incorporated into the fifth and sixth editions of the APA Publication Manual, by calling CIs "in general, the best reporting strategy" (APA, 2001, p. 22; APA, 2009, p. 34). Earlier, between 1994 and 1997, as editor of *Memory & Cognition*, Geoffrey Loftus had tried to reform the publication practices of the journal. He encouraged the use of error bars and avoidance of NHST. Although there was a temporary effect of his policy, it seemed hardly effective in the long run (Finch et al., 2004).

The argument for the use of CIs, that they are accessible and comprehensible, rests on the idea that researchers can properly interpret them; including information that researchers cannot interpret is, at best, of limited use and, at worst, potentially misleading. Several previous studies have explored whether presenting results using CIs leads to better interpretations than does

presenting the same results using NHST. Belia, Fidler, Williams, and Cumming (2005) showed that there was a lack of knowledge of participants about the relationship between CIs and significance levels, suggesting that when data are presented with CIs versus significance test outcomes, people might interpret the same results differently. Fidler and Loftus (2009) found that both first-year and masters students tended to make the mistake of accepting the null hypothesis more often if data were presented using NHST than if the same data were presented using CIs. Hoekstra, Johnson, and Kiers (2012) found similar effects for researchers and also found that presenting the same data by means of NHST or CIs affected researchers' intuitive estimates about the certainty of the existence of a population effect or the replicability of that effect.

The findings above show that people interpret data differently depending on whether these data are presented through NHST or CIs. Despite the fact that CIs are endorsed in many articles, however, little is known about how CIs are generally understood by researchers. Fidler (2005) showed that students frequently overlooked the inferential nature of a CI and, instead, interpreted the CI as a descriptive statistic (e.g., the range of the mid C % of the sample scores, or even an estimate for the sample mean). Hoenig and Heisey (2001) stated that it is "surely prevalent that researchers interpret confidence intervals as if they were Bayesian credibility regions" (p. 5), but they did this without referring to data to back up this claim.

Before proceeding, it is important to recall the correct definition of a CI. A CI is a numerical interval constructed around the estimate of a parameter. Such an interval does not, however, directly indicate a property of the parameter; instead, it indicates a property of the procedure, as is typical for a frequentist technique. Specifically, we may find that a particular procedure, when used repeatedly across a series of hypothetical data sets (i.e., the sample space), yields intervals that contain the true parameter value in 95 % of the cases. When such a procedure is applied to a particular data set, the resulting interval is said to be a 95 % CI. The key point is that the CIs do not provide for a statement about the parameter as it relates to the particular sample at hand; instead, they provide for a statement about the performance of the procedure of drawing such intervals in repeated use. Hence, it is incorrect to interpret a CI as the probability that the true value is within the interval (e.g., Berger & Wolpert, 1988). As is the case with *p*-values, CIs do not allow one to make probability statements about parameters or hypotheses.

In this article, we address two major questions: first, the extent to which CIs are misinterpreted by researchers and students, and second, to what extent any misinterpretations are reduced by experience in research. To address these questions, we surveyed students and active researchers about their interpretations of CIs.

Method

Participants and procedure

Our sample consisted of 442 bachelor students, 34 master students, and 120 researchers (i.e., PhD students and faculty). The bachelor students were first-year psychology students attending an introductory statistics class at the University of Amsterdam. These students had not yet taken any class on inferential statistics as part of their studies. The master students were completing a degree in psychology at the University of Amsterdam and, as such, had received a substantial amount of education on statistical inference in the previous 3 years. The researchers came from the universities of Groningen (n = 49), Amsterdam (n = 44), and Tilburg (n = 27).

Participants were given the paper survey to complete, and they were instructed not to use outside information when giving their answers. Bachelor and master students were asked to complete the questionnaire individually during a lecture. To motivate the first year students, we raffled five times 50 Euros among those who answered all six knowledge questions correctly. In Groningen and Amsterdam, we went door to door in the psychology departments, visiting the researchers in their offices and asking them to complete the questionnaire; in Tilburg, the researchers completed the questionnaire in advance of a presentation by the last author. Potential participants were told that the data would be processed anonymously but that their names could be noted on a separate list in case they would like to have more information about the study afterward. The willingness to participate was almost unanimous, with only a few people refusing because of other activities.

Materials

The questionnaire, which can be found in Appendix 2, opened with a fictitious scenario of a professor who conducts an experiment and reports a 95 % CI for the mean that ranges from 0.1 to 0.4. Neither the topic of study nor the underlying statistical model used to compute the CI was specified in the survey. Subsequently, six statements representing possible misconceptions of the CI were presented, for which the participants had to indicate whether they considered this statement true or false. "False" was defined as a statement that does not follow logically from the professor's result, and it was explicitly noted that all, several, or none of the statements could be correct. As can be seen from Appendices 1 and 2, the wording and structure of our CI questionnaire were designed

to mimic the *p*-value questionnaire as presented by Gigerenzer (2004), who adopted this questionnaire from Oakes (1986) and Haller and Krauss (2002). Participants could indicate the correctness of a statement by checking one of two boxes (labeled "true" and "false") next to the statement. At the end of the questionnaire, participants were asked to rate their own knowledge of statistics on a scale ranging from 1 (i.e., *no stats courses taken, no practical experience*) to 10 (i.e., *teaching statistics at a university*).

The questionnaire featured six statements, all of which were incorrect. This design choice was inspired by the *p*-value questionnaire from Gigerenzer (2004). Researchers who are aware of the correct interpretation of a CI should have no difficulty checking all "false" boxes. The (incorrect) statements are the following:

1. The probability that the true mean is greater than 0 is at least 95 %.
2. The probability that the true mean equals 0 is smaller than 5 %.
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect.
4. There is a 95 % probability that the true mean lies between 0.1 and 0.4.
5. We can be 95 % confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4.

Statements 1, 2, 3, and 4 assign probabilities to parameters or hypotheses, something that is not allowed within the frequentist framework. Statements 5 and 6 mention the boundaries of the CI (i.e., 0.1 and 0.4), whereas, as was stated above, a CI can be used to evaluate only the procedure and not a specific interval. The correct statement, which was absent from the list, is the following: "If we were to repeat the experiment over and over, then 95 % of the time the confidence intervals contain the true mean."

Results

Of 3,372 total responses (six questions times 562 participants), one response was ticked as both "true" and "false," and one other response was missing (that is, neither "true" nor "false" was checked). All data of these 2 participants were removed from the analysis.

Items endorsed

Table 1 shows the percentages of the different groups of participants endorsing each of the six statements. Since all statements were incorrect, the number of items answered incorrectly equals the number of items endorsed. The mean numbers of items endorsed for first-year students, master students, and researchers were 3.51 (99 % CI = [3.35, 3.68]), 3.24 (99 % CI = [2.40, 4.07]), and 3.45 (99 % CI = [3.08, 3.82]), respectively. The item endorsement proportions are presented per group in Fig. 1. Notably, despite the first-year students' complete lack of education on statistical inference, they clearly do not form an outlying group.

Table 1 Percentages of students and teachers endorsing an item

Statement	First Years ($n = 442$)	Master Students ($n = 34$)	Researchers ($n = 118$)
The probability that the true mean is greater than 0 is at least 95 %	51 %	32 %	38 %
The probability that the true mean equals 0 is smaller than 5 %	55 %	44 %	47 %
The "null hypothesis" that the true mean equals 0 is likely to be incorrect	73 %	68 %	86 %
There is a 95 % probability that the true mean lies between 0.1 and 0.4	58 %	50 %	59 %
We can be 95 % confident that the true mean lies between 0.1 and 0.4	49 %	50 %	55 %
If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4	66 %	79 %	58 %

The distributions of the number of items that the different groups endorsed are presented in Table 2. With the exception of the master students, the distributions of endorsed responses are unimodal, and none of the three groups yielded a skewed

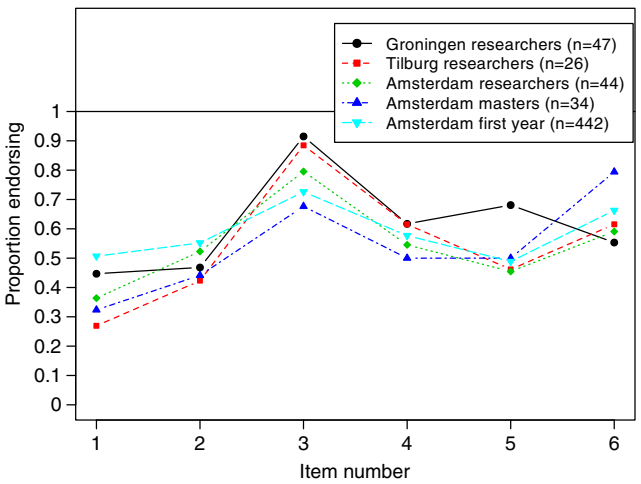


Fig. 1 The proportions of participants endorsing a certain item, separate for each expert group. The response patterns are highly similar, regardless of the difference in statistical expertise between the groups. Note that the correct proportion of endorsed items is 0

Table 2 Percentages of the number of items that students and teachers endorsed (NB. All statements were false and none of the items should be endorsed)

Number	First-Year Students (n = 442)	Master Students (n = 34)	Researchers (n = 118)
0	2 %	0 %	3 %
1	6 %	24 %	9 %
2	14 %	18 %	14 %
3	26 %	15 %	25 %
4	30 %	12 %	22 %
5	15 %	21 %	16 %
6	7 %	12 %	11 %

distribution, indicating that means were not excessively influenced by a few participants with a high number of incorrect answers. The data reveal that the level of knowledge throughout the sample was consistently low. Only 8 first-year students (2 %), no master students, and 3 postmasters researchers (3 %) correctly indicated that all statements were wrong.

Experience

In Fig. 2, the participants' self-indicated level of experience on a scale from 0 to 10 is shown in relation to the number of items they endorsed. Figure 2 does not indicate that more experienced researchers endorse fewer items. Indeed, the correlation between endorsed items and experience was even slightly positive (0.04; 99 % CI = [−0.20; 0.27]), contrary to what one would expect if experience decreased the number of misinterpretations.

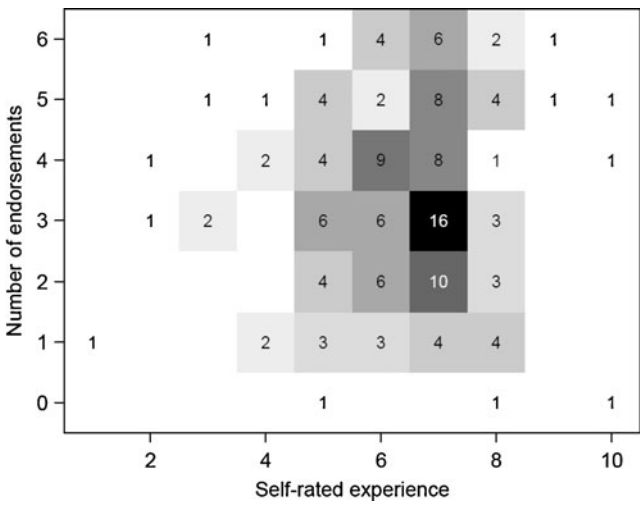


Fig. 2 The number of endorsements for the six items given the indicated level of experience with statistics on a scale of 1 to 10. Darker dots indicate more participants with the same level of indicated experience endorsing the same number of items, and the exact number is indicated by the number within these squares

Conclusion and discussion

CIs are often advertised as a method of inference that is superior to NHST (e.g., Cumming & Finch, 2001; Fidler & Loftus, 2009; Schmidt, 1996; Schmidt & Hunter, 1997). Moreover, the APA Manual (APA, 2009) strongly encourages the use of CIs. Consequently, one might have expected that most researchers in psychology would be well-informed about the interpretation of this rather basic inferential outcome. Our data, however, suggest that the opposite is true: Both researchers and students in psychology have no reliable knowledge about the correct interpretation of CIs. Surprisingly, researchers' self-reported experience in statistics did not predict the number of incorrectly endorsed interpretations. Even worse, researchers scored about as well as first-year students without any training in statistics, suggesting that statistical education and experience in conducting research do not prevent the intuitive misinterpretations that are invited by the term *confidence interval*.

Our results are consistent with previous studies suggesting that some misinterpretations were to be expected. In a study by Kalinowski (2010), for example, a quarter of the graduate students had difficulties with the relationship between the confidence level and the width of an interval, and a third of the students believed that the "subjective likelihood distribution" (Kalinowski's term for a posterior under a noninformative prior) underlying a CI was uniform rather than bell-shaped. Fidler and Loftus (2009) found that a CI including the value of H_0 was incorrectly interpreted as "no effect" by about 30 % of the students, and Fidler (2005) showed that students misinterpreted a CI as the range of observed scores, the range of scores within one standard deviation, or the range of plausible values for the sample mean in more than half of the cases. These previous studies, however, showed misunderstandings among students; here, we show dramatic and similar levels of misinterpretation among both researchers and students.

No clear pattern is apparent in researchers' misinterpretations of CIs. Although, for example, Hoenig and Heisey (2001) predicted that a Bayesian interpretation of CIs would be prevalent, the items that can clearly be considered Bayesian statements (1–4) do not seem to be preferred over item 6, which is clearly a frequentist statement. The only item that seems to stand out is item 3, which makes a statement about the likelihood of the null hypothesis being true (note that a CI does not require a null hypothesis in the first place), suggesting that CIs are merely seen as an alternative way of conducting a significance test, even by people who have no experience with CIs or significance tests.

Although the full impact of these misunderstandings of CIs is not known, it seems clear that if statistical tools play a role in the justification of knowledge in the social sciences, researchers using these tools must understand them. One might

argue, however, that in many cases, the numbers making up the CI endpoints might have several interpretations: for instance, frequentist CIs can be numerically identical to Bayesian credible intervals, if particular prior distributions are used (Lindley, 1965). If this is true, one could take the position that any one of several interpretations would do. We believe this position to be mistaken, for two reasons.

First, treating frequentist and Bayesian intervals as interchangeable is ill-advised and leads to bad "Bayesian" thinking. Consider, for instance, the frequentist logic of rejecting a parameter value if it is outside a frequentist CI. This is a valid frequentist procedure with well-defined error rates within a frequentist decision-theoretic framework. However, some Bayesians have adopted the same logic (e.g., Kruschke, Aguinis, & Joo, 2012; Lindley, 1965): They reject a value as not credible if the value is outside a Bayesian credible interval. There are two problems with this approach. First, it is not a valid Bayesian technique; it has no justification from Bayes's theorem (Berger, 2006). Second, it relies on so-called "noninformative" priors, which are not valid prior distributions. There are no valid Bayesian prior distributions that will yield correspondence with frequentist CIs (except in special cases), and thus inferences resulting from treating CIs as credible intervals must be incoherent. Confusing a CI for a Bayesian interval leaves out a critically important part of Bayesian analysis—choosing a prior—and, as a result, leaves us with a non-Bayesian technique that researchers believe is Bayesian.

Second, a lack of understanding of frequentist concepts underlies many common questionable research practices; misinterpretations of frequentist statistics are inextricably linked with one another. To give two examples, *p*-value snooping (Stone, 1969) occurs when researchers misinterpret the *p*-value as a cumulative measure of evidence, forgetting its long-run distribution under the null hypothesis, which is different when the stopping rule depends on the *p*-value. Likewise, failures to account for multiple comparisons in *p*-values and CIs (Curran-Everett, 2000) arise from interpreting the *p*-value as a measure of evidence or the CI as a set of "plausible" values, respectively, while misunderstanding the importance of family-wise probabilities. In these two examples, understanding the interpretation of the statistic helps to define the limitations of the use of the procedure. If CIs are to become more commonly used statistics, it is critically important that misinterpretations of CIs be avoided.

One could question whether our findings indicate a serious problem in scientific practice, rather than merely an academic issue. We argue that they do indicate a serious problem, one closely related to what some authors refer to as a crisis in in the social and behavioral sciences (e.g., Pashler & Wagenmakers, 2012). This crisis has several components: Many results are irreproducible (Nosek, Spies, & Motyl, 2012), publication bias seems to be pervasive (Morey, 2013), and current methods such as NHST are fundamentally flawed (Schmidt, 1996). At its core, this crisis is about the justification of knowledge. Researchers in the social and behavioral social sciences make claims that are simply unsupported by the methods they use. That there is a crisis would seem hardly surprising, given that the basic tools of the trade—*p*-values and CIs—are thought to be something they are not. It has been argued that CIs are preferable to NHST (e.g., Cumming & Fidler, 2009; Schmidt, 1996), since they prevent some misinterpretations that would be found with NHST (Fidler & Loftus, 2009; Hoekstra et al., 2012). This is a decidedly low bar to set given the extreme number and severity of the misunderstanding of NHST (even setting aside the perniciousness of NHST as a method). Given the opportunity to abandon NHST, we should compare CIs against possible replacements, not against NHST. Against other methods, CIs fall short and can behave in strange ways (the reasons for this are beyond the scope of this article; see, e.g., Blaker & Spjøtvoll, 2000; Jaynes, 1976).

Acknowledgements This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council, and by National Science Foundation Grants BCS-1240359 and SES-102408.

Appendix 1 Questionnaire on p-values (Gigerenzer, 2004)

(The scenario and the table are reproduced verbatim from Gigerenzer [2004, p. 594].)

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means *t*-test and your result is significant ($t = 2.7$, d.f. = 18, $p = 0.01$). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct (between the population means).

1. You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).

[] true/false []

2. You have found the probability of the null hypothesis being true.

[] true/false []

3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

[] true/false []

4. You can deduce the probability of the experimental hypothesis being true.

[] true/false []

5.You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.

[] true/false []

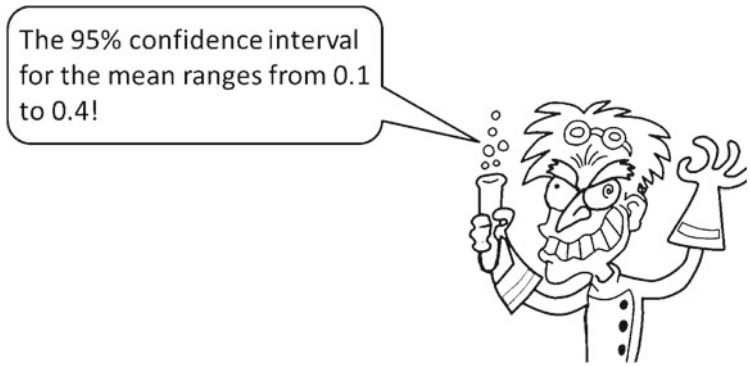
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99 % of occasions.

[] true/false []

Appendix 2 Questionnaire on confidence intervals

(The questionnaires for the students were in Dutch, and the researchers could choose between an English and a Dutch version.)

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:



Please mark each of the statements below as "true" or "false". False means that the statement does not follow logically from Bumbledorf's result. Also note that all, several, or none of the statements may be correct:

1. The probability that the true mean is greater than 0 is at least 95%.	☐True	☐False
2. The probability that the true mean equals 0 is smaller than 5%.	☐True	☐False
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect.	☐True	☐False
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.	☐True	☐False
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.	☐True	☐False
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.	☐True	☐False

Please indicate the level of your statistical experience from 1 (no stats courses taken, no practical experience) to 10 (teaching statistics at a university): _____

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

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

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1,* 385–402.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics.

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

Blaker, H., & Spjøtvoll, E. (2000). Paradoxes and improvements in interval estimation. *The American Statistician, 54,* 242–247.

Chow, S. L. (1998). A précis of "Statistical significance: Rationale, validity and utility. *Behavioral and Brain Sciences, 21,* 169–194.

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

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2,* 161–172.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and non-central distributions. *Educational and Psychological Measurement, 61,* 532–574.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie, 217*, 15–26. doi:10.1027/0044-3409.217.1.15

Curran-Everett, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 279,* R1–R8.

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

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5,* 75–98.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology.* Unpublished doctoral dissertation, University of Melbourne, Melbourne.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217,* 27–37.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of memory and cognition. *Behavior Research Methods, Instruments, & Computers, 36,* 312–324.

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

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? Methods of Psychological Research Online [On-line serial], 7, 120. Retrieved May 27, 2013, from www2.uni-jena.de/svw/metheval/lehre/0405-ws/evaluationuebung/haller.pdf

Harlow, Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p-values. *Psychonomic Bulletin & Review, 13,* 1033–1037.

Hoekstra, R., Johnson, A., & Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement, 72,* 1039–1052.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (pp. 175–257). Dordrecht, The Netherlands: Reidel Publishing Company.

Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. Proceedings of the Eighth International Conference on Teaching Statistics. [CDROM]. IASE, Lijbljana, Slovenia, Refereed paper.

Kline, R. B. (2004). *Beyond significance testing: reforming data analysis methods in behavioral research.* Washington DC, USA: American Psychological Association.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15,* 722–752. doi:10.1177/1094428112457829

Lecoutre, M.-P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis tests. *International Journal of Psychology, 38,* 37–45.

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint. Part 2.* Cambridge: Inference. Cambridge University Press.

Morey, R. D. (2013). The consistency test does not-and cannot-deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology.* doi:10.1016/j.jmp.2013.03.004

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631. doi:10.1177/1745691612459058

O'Hagan, A. (2004). Dicing with the unknown. *Significance, 1,* 132–133.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chicester: John Wiley & Sons.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Scheutz, F., Andersen, B., & Wulff, H. R. (1988). What do dentists know about statistics? *Scandinavian Journal of Dental Research, 96,* 281–287.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Sellke, T., Bayarri, M.-J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician, 55,* 62–71.

Stone, M. (1969). The role of significance testing: Some data with a message. *Biometrika, 56,* 485–493.

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

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist, 4,* 140–143.

Wulff, H. R., Andersen, B., Brandenhoff, P., & Guttler, F. (1987). What do doctors know about statistics? *Statistics in Medicine, 6,* 3–10.