The use and usability of inferential techniques
Hoekstra, Rink

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 29-04-2019
6. The Influence of Presentation on the Interpretation of Inferential Results

Abstract
CIs have frequently been presented as a useful addition to or alternative for NHST. In the fifth edition of the APA manual it is stated that CIs are usually “the best reporting strategy”. Fidler (2005) showed that the frequency of misinterpretations for results presented by means of CIs are lower than those for the same data presented by means of NHST outcomes. Little is known, however, about whether the subjective estimates that arguably play an implicit role in most interpretations of scientific results are different for the same results presented by means of CIs or NHST outcomes. In the present study, 71 participants were asked to interpret outcomes of fictitious studies. We found that for significant outcomes, participants tend to be more certain about the existence of a population effect in the expected direction and about replicability of their results when the results are presented by means of NHST than when presented by means of CIs, whereas such a difference could not be found for clearly non-significant findings. The result of Fidler of more misconceptions for NHST presented results was also confirmed. Apparently, a significant finding is more convincing for researchers than the same effect presented by means of a CI.

6.1 Introduction
As mentioned in Chapter 1, NHST has been heavily criticized for decades for not being a useful inferential technique. CIs should offer a solution for many problems associated with the use of NHST. In addition to combining information on location and precision, CIs convey information on effect size (Schmidt, 1996), and should thus reduce binary thinking (Wilkinson & Task Force on Statistical Inference, 1999). We define binary
thinking in the NHST context as an exaggerated focus on whether or not the result is significant, with little or no attention for the size of the effect. Furthermore, it has been argued that CIs make meta-analysis easier and less problematic (Cumming & Finch, 2001).

The statistical relation between NHST and CIs can lead people to interpret CIs as if they were significance tests (Harris, 1997). When a CI is only used to check whether the null value lies within the interval, the procedure is no different from checking whether a research outcome reaches significance or not. Therefore, using CIs does not prelude binary judgements. As Abelson (1997), one of the members of the Task Force on Statistical Inference, put it: “Under the Law of the Diffusion of Idiocy, every foolish application of significance testing is sooner or later going to be translated into a corresponding foolish practice for confidence limits” (p. 130). CIs are, however, completely different on a psychological level (Feynmann, 1967). On this psychological level, Schmidt and Hunter (1997) argued, a CI is easier to interpret, because it is a visual representation of effect size and measure of uncertainty, and thus both can be seen at a single glance. To interpret $p$-values combined with a standardised measure of effect size, on the other hand, both statistics need to be combined mentally, and, according to Schmidt and Hunter, this is harder to do.

The question now is: Can the problems with NHST be overcome if CIs are used to replace NHST results? Although many authors have suggested the use of CIs as an alternative to NHST, strikingly few studies have investigated whether the interpretation of data would be different when $p$-values would be replaced by results in terms of CIs.

Fidler (2005) is one of the few to have investigated whether fewer errors of interpretation are made when results are presented by means of CIs rather than by means of NHST. Specifically, she focussed on the mistake of accepting the null-hypothesis. This mistake is made when a non-significant effect is interpreted as proof for the absence of an effect in the population.
The same mistake is made when a CI including the null value is seen as proof for the absence of an effect. In this study, participants were given scenarios with inferential outcomes for non-significant results (presented by means of CIs or $p$-values), and were asked to indicate their belief of the existence of an effect on a scale with five possible answers. Two of those answers indicated different degrees of belief in the absence of an effect, and whenever one of these was selected, this was regarded as an occurrence of the mistake of accepting the null hypothesis.

Fidler (2005) found, using a sample of students, that the proportion of occurrences of accepting the null-hypothesis was smaller when CIs were presented than when results were presented following the logic of NHST, with 39% (95%CI = [29, 50]) of the participants accepting the null hypothesis at least once out of two occasions in the NHST condition, and 13% (95%CI = [5, 29]) in the CI condition. Although a small proportion of the participants who did not show this error in the NHST condition did show this error in the CI condition, the results seem to indicate that, in general, presenting results in the form of CIs decreases the proportion of occurrences of accepting the null hypothesis compared to the same results presented by means of $p$-values.

In a follow-up study, Fidler (2005) showed that the interpretation of CIs is not without problems either. When asked to interpret a CI, many participants seemed to deem the CI a descriptive statistic, rather than an inferential statistic. Only 20% (95%CI = [11, 34]) of the participants made a statement that showed awareness of the inferential nature of CIs. That is, they recognized the fact that the main goal of using CIs is to draw conclusions about the population instead of the sample.

Since Fidler (2005) used students as participants, it is unclear how researchers in psychology interpret NHST differently from CIs. In this Chapter, we will focus on two measures for measuring differences in interpretations: The degree of belief that an effect exists in the population...
(hereafter referred to as certainty), and the degree of belief that a replication study will result in a significant effect (hereafter referred to as replicability). We selected certainty as a measure of interpretation because we think that usually it is certainty that a researcher is interested in when applying inferential statistics. When the complete population is not known (which is the only case in which inference is necessary), certainty cannot be quantified. In scientific practice, replication studies should be conducted to approximate certainty. Replicability can thus be considered an indirect measure of certainty.

As far as certainty is concerned, it seems probable that people will be more aware of uncertainty in light of a significant effect when it is presented by means of a CI instead of a $p$-value, resulting in a lower estimate for certainty in the CI case. When an effect is not significant, people will likely be less certain of an effect when this effect is accompanied by a $p$-value compared to when the same data are presented by means of a CI. We do not know of any study comparing this type of interpretation of CIs and NHST outcomes.

For our study, we have the following hypotheses: (1) Results presented by means of CIs are interpreted differently than the same results presented by means of NHST. (2) The significance of effects interacts with this relationship between way of presenting outcomes and the interpretation of those outcomes, and (3) these differences can be explained by unfamiliarity with confidence intervals and misconceptions about NHST. Furthermore, we will study the relative subjective importance of different statistical outcomes. For the first hypothesis, we were interested in three sub-questions: Do interpretational errors occur less frequently in CI presented results? Are certainty estimates lower for CI presented results compared to NHST presented results? And are replicability estimates lower for CI presented results compared to NHST presented results?
6.2 Method

Participants

For this study, 71 Dutch speaking Ph.D. students working at different psychology departments were used as participants, 45 of them were women and 26 were men, all aged between 25 and 32. All had at least one year of experience in doing research, and none of them participated in one of the previous studies described in this thesis. The participants were selected from the University of Amsterdam, the Free University of Amsterdam, the University of Twente, the University of Groningen, the University of Leiden, the Erasmus University, Tilburg University, and Maastricht University. Prospective participants were first approached by e-mail, and called afterwards to ask whether they were willing to participate. If they were willing to participate, a file with a programme leading the participant through a series of judging tasks was sent, and they were asked to perform the tasks within two weeks.

Of the 149 participants who were approached, 71 actually performed the tasks. The participation was voluntary. Among the participants five gift certificates with a value of 20 Euros were raffled.

Tasks

The tasks were offered by a programme in Delphi, and jointly took approximately half an hour. The tasks consisted of judging eight scenarios in random order. In these scenarios, statistical results of fictitious medical studies were presented. In all scenarios, the effects of a certain new medicine were compared to the effects of a placebo by means of difference scores. Four of the scenarios were presented by means of a CI (see left panel Figure 6.1), and four were presented by means of effect size combined with significance test outcomes (see right panel Figure 6.1).
The Influence of Presentation on Interpretation of Results

Figure 6.1: Screenshots of results as presented in the experiment. In the left panel the results are shown by means of a 95% CI, the right panel shows the results presented by means of NHST. “Toename” means increase, and “Standaardfout” means standard error in Dutch. Note that the effect presented in the right panel is the same effect as presented in the left panel, multiplied by four, to prevent participants from recognising results.

In both cases, the effect size was presented in a figure, and sample size and standard error were given separately. The significance test outcomes were presented by means of a $t$-value with the degrees of freedom, and a $p$-value. The results in every CI presented scenario were identical to one of the NHST presented scenarios, except for a multiplication factor. This multiplication factor was introduced to prevent participants from recognizing previously seen results.

Our second research question was about the influence of "degree of significance" (as operationalised by means of the $p$-value) on the way CI results and NHST are interpreted. To study this, we varied the degree of
significance of the results by presenting four non-significant and four significant scenarios, equally divided among both presentation conditions. The corresponding $p$-values were .02, .04, .06 and .13.

For studying to what extent possible differences could be explained by unfamiliarity with confidence intervals (our third research question), two versions of the tasks were made. The versions were randomly assigned to the participants. In one version, an instruction about the interpretation of CIs and NHST was given, in the second this was not. In both versions, the tasks started with a short multiple choice test with four questions about the interpretation of CIs, and a similar test was given after completion of the tasks. In this way, effects of the instruction could be measured.

Finally, we were interested in the subjective importance of the different statistical outcomes used in this experiment. We expected people to attach great importance to $p$-values as opposed to CIs, despite the fact that CIs incorporate the same information and more than $p$-values. To study this relative importance, the participants were also asked to order standard error, t-value, sample size, confidence interval, $p$-value, and mean with respect to their subjective importance.

**Scoring**

For studying whether interpretational errors occur less frequently in CI presented results compared to NHST outcomes, the subjects were asked to select on the following five-point Likert scale the statement they regarded most true:

1. “There is strong support for the existence of an effect”,
2. “There is moderate support for the existence of an effect”,
3. “The results do not clearly suggest the existence or absence of an effect”,
4. “There is moderate support for the absence of an effect”, or
5. “There is strong support for the absence of an effect”.

The latter two were regarded as incorrect statements, because results as given in the scenarios can never be seen as proof for the absence of an effect. In this study, we counted the frequency of subjects selecting one of the latter two answers of the Likert scale, and regarded these as a measure for the frequency of the mistake of accepting the null hypothesis (H₀). This way of searching for occurrences of the error of accepting the null hypothesis was similar to that used in the study by Fidler (2005). Furthermore, we asked the subjects to answer the following open question after every trial: “What do these results tell you about the situation in the population?” The answers were checked for the occurrences of four types of task behaviour. Phrases like “there is no effect” or “the absence of an effect” were interpreted as the error of accepting the null hypothesis. The mistake of accepting the alternative hypothesis (Hₐ), without taking uncertainty into account was defined as interpreting a significant effect as proof for the existence of an effect (e.g., by stating that “there is an effect in the population”). Reference to effect size was coded whenever the sample mean was mentioned or interpreted. Reference to significance was coded whenever the terms “significant” or “non-significant” were mentioned.

For studying whether certainty and replicability estimates differed for CI and NHST presented results, we asked the subjects to estimate both probabilities in percentages on a scale of 0 to 100. Only integers could be entered. For certainty estimates, the question was: “How large do you estimate the probability that there is an effect in the expected direction in the population based on these results?” For replicability, we asked: “How large do you estimate the probability that you would find a significant effect if you would do the same study again?”. It was stressed in the introduction that exact answers could not be calculated.
6.3 Results

We expected that for results presented as CIs we would find more occurrences of accepting $H_0$ and accepting the alternative hypothesis $H_a$, more references to effect size and fewer references to significance, compared to results presented by NHST outcomes. The results (see Table 6.1) confirm these expectations.

The two ways of measuring accepting $H_0$ both indicate that accepting $H_0$ occurs more frequently in the NHST conditions, but the size of the effect is different. Apparently, using a five point Likert scale is a more lenient method to show this effect. Given the lenient method, the results for the mean difference between the NHST and the CI conditions show that the effect is probably at least 10%, and therefore substantial. For accepting $H_0$ and $H_a$ based on the open questions, such a clear difference cannot be seen (see column on the right in Table 6.1). Furthermore, there seem to be more references to effect size and fewer to significance in the CI condition, which is supported by the bounds of the CIs around the mean differences between the two conditions. It should be noted here that the percentages of observed task behaviour are relatively small in all cases.

Table 6.1: Occurrences of task behaviour for NHST and CI presented results in percentages. The percentages of the first four behaviour types were from the written conclusions, whereas the last were from the multiple choice question. Between square brackets 95% CIs are given.

<table>
<thead>
<tr>
<th></th>
<th>NHST</th>
<th>CI</th>
<th>(NHST-CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepting $H_0$ (five-point scale question)</td>
<td>47 [39,55]</td>
<td>26 [20,38]</td>
<td>20 [10,32]</td>
</tr>
<tr>
<td>Accepting $H_0$ (open questions)</td>
<td>12 [7,18]</td>
<td>7 [3,12]</td>
<td>5 [-2,12]</td>
</tr>
<tr>
<td>Accepting the $H_a$</td>
<td>6 [3,12]</td>
<td>2 [1,6]</td>
<td>4 [-1,9]</td>
</tr>
<tr>
<td>Reference to effect size</td>
<td>39 [33,44]</td>
<td>54 [48,60]</td>
<td>-15 [-7,-23]</td>
</tr>
<tr>
<td>Reference to significance</td>
<td>59 [53,64]</td>
<td>34 [29,40]</td>
<td>25 [17,34]</td>
</tr>
</tbody>
</table>
The mean certainty and replicability estimates made in the NHST and CIs conditions are shown in Figure 6.2. Estimates in the NHST conditions are on average higher than estimates for the same in the CI conditions. On average, the certainty estimates for the CI conditions are 9.1 points lower than for NHST conditions (95% CI = [2.4, 15.8]).

Figure 6.2: Mean probability estimates for different p-values in NHST conditions and CI conditions for certainty (left panel) and replicability (right panel). The error bars indicate 95% CIs around the means.

Replicability estimates for CI conditions were on average 6.1 points lower (95% CI = [0.2, 12.0]). This supports our expectation that CIs make people more aware of the uncertainty that is inextricably associated with every inferential interpretation. It is not just the presentation (CI or NHST) that seems to account for this difference. For certainty, and to a lesser extent for replicability, the degree of significance of the results also seems to influence
the relationship between presentation and the average estimates. Whereas for
significant or almost significant results ($p=.02, p=.04, p=.06$) there are clear
differences between average estimates for both presentation ways, for the
clearly non-significant results ($p=.13$) differences virtually disappeared.

The subjects were also asked to rank the statistical outcomes on
importance for their interpretations. Given six statistics, each could get a
rank number from 1 to 6. The average rank for the $p$-value was 2.8 (95%CI =
[2.4, 3.2]), whereas the average rank for CIs was 3.7 (95% CI = [3.3, 4.1]).
The mean difference of 0.9 in rank (95%CI = [0.7, 1.1]) confirms our
expectation that CIs are in general considered less important than $p$-values,
despite the fact that they are more informative. Means are considered
relatively important as well, with mean rank 3.3 (95%CI = [2.9, 3.7]). The
mean ranks of sample size, the t-value and standard error were respectively
3.2, 4.1 and 4.1.

Approximately half of the subjects received an instruction
containing brief explanations of how to interpret NHST and CIs. It was
expected that subjects who received an instruction would make fewer
interpretational mistakes. This instruction, however, seemed to have had
little or no effect. The subjects who received such an instruction did not refer
more frequently to effect size, and did not make fewer interpretational
mistakes than subjects who did not receive instructions (the percentage of
mistakes was even slightly higher for the subjects with instruction than for
those without).

6.4 Discussion

The presentation of inferential results seems to influence the
interpretation of these results of researchers in psychology. It was found that
mistakes were made more frequently when subjects were confronted with
NHST outcomes compared to CIs (confirming what Fidler, 2005, found),
and that in the CI condition more frequent references to effect size were
made. Furthermore, we found that researchers seem more certain of significant or marginally significant results when they are presented by means of NHST instead of CIs. We did not find such a clear difference for non-significant findings, and these effects were stronger for certainty than for replicability.

The results suggest that when the results are presented by means of CIs rather than by NHST, there is more attention for effect size, and the estimates for significant effects are more conservative. This seems to indicate that a good practice of inference is better warranted by CIs, but this might be premature. More conservative estimates are not necessarily better estimates. Only when assuming that researchers are, in general, too confident about significant effects, which is implicitly suggested in the discussion on the binary interpretation of NHST (e.g., Rosnow & Rosenthal, 1989), more conservative estimates might indicate a better research practice. The results also clearly show that even when data are presented by means of CIs, referring to effect size is far from standard, and interpretational mistakes are still made relatively frequently, although less often than when the data are presented by means of NHST outcomes. For that reason, replacing all NHST results by means of CIs does not seem an answer to all problems. Furthermore, as Fidler (2005) also found, CIs can be misinterpreted as well. Many conclusions contained phrases clearly expressing misunderstanding of the inferential nature of CIs (e.g., interpreting the CI as a descriptive statistic).

We expected that some of those problems might be attributed to lack of understanding of CIs. For that reason, we expected that instructions explaining how to interpret both NHST and CI presented results would decrease the frequency of mistakes in the instruction condition compared to the condition in which no instructions were given, but these instructions did not show the expected effects. The absence of clear effects might be explained by the fact that, despite explicit statements to read the instruction
thoroughly, it is known that subjects are rather careless about reading instructions, and prefer to start the task as soon as possible. Another explanation could be the fact that if researchers do not know how to interpret CIs in the first place, reading a short instruction might not be enough to get a clear understanding of the concept of CIs.

The fact that estimates of certainty and replicability for significant or almost significant results presented by CIs are lower than when these results are presented by NHST, might also be due to unfamiliarity with CIs (which would be a logical explanation for a lower estimate). If this would be the case, however, it would seem logical to find this also for non-significant findings, and this did not seem to be the case. Therefore, the results could suggest that CI presented results indeed make researchers more aware of uncertainty.

The ranking of the different statistical outcomes confirms the expectation that the \( p \)-value is given more subjective importance than means and than CIs, despite the fact that the latter combines information of the first two outcomes. Although this result may not be surprising, it is notable that the average rank of CIs was below average (3.5), given six outcomes. It has to be studied to what extent this can be explained by a lack of popularity of CIs, or by researchers’ lack of experience with them.

In summary, adding CIs to results, or even replacing NHST results by CIs, might improve the practice of inference, but it must certainly not be seen as a panacea for all interpretational pitfalls. It goes without saying that, no matter which inferential technique is used for the presentation of research outcomes, logical reasoning and awareness of the essentials of inference should always remain the main focus of every researcher.