**Supplementary analyses and results for:**

**Beyond Psychology: Prevalence of *p*-value and confidence interval misinterpretation across different fields**

**Supplementary analysis and result 1: Misinterpretation rate among postgraduates and researchers**

 Considering of the fact that some Bachelor degree students may have no experience in research, we analyzed the data of postgraduates and researchers (*N* = 1129). The results showed that 89% postgraduates or researchers had at least one wrong answer of *p*-value, and 93% for CIs. For detailed information of different educational background, see **Fig. S1 a & b.**



**Fig. S1.** Percentage of misinterpretation of *p*-value and CIs (speared items)

**Supplementary analysis and result 2: non-significant condition vs. significant condition**

 In previous surveys, researchers were only asked to interpret the significant results (i.e., *p* < .05, and CIs > 0). In our survey, we also included a condition in which the *p-value* was not significant (*t*(98) = 1.26, *p* = 0.21). Our results revealed that the error rate under nonsignificant *p-value* condition (*N* = 759, 92% had at least one wrong answer) was smaller than that under significant *p-value* condition (*N* = 720, 86% had at least one wrong answer), χ2 (1) = 16.841, *p* <.001, BF10 = 543.871 (See **Fig. S2 a**).

However, for the interpreting the results of CIs (See **Fig. S2 b**), we didn’t find evidence for the differences between respondents who read 95% CIs that included zero (91% had at least one wrong answer) and respondents who read CIs didn’t include zero (94% had at least one wrong answer), χ2 (1) = 2.892, *p* = .049, BF10 = 0.580. We further performed a *z*-test on the difference of difference between the error rates in the significant and nonsignificant condition under *p*-value and CI scenario to test the interaction effect of *significance* (significant *vs* nonsignificant) and statistical indices (*p*-value *vs* CI). Results showed that the interaction effect did exist, *z* = 4.24, *p* < .001.



**Fig. S2.** Misinterpretation of *p-value* in significant vs. non-significant results

**Supplementary analysis and result 3: Sub-group analysis for the difference between mainland China and Overseas**

Given that there are large proportion of respondents with an abroad highest degree are from statistics or mathematics background, it might be that the differences between mainland China respondents and overseas respondents were driven by respondents from specific field. Therefore, we further compared the two sub-groups of them.

For those who major in Math and Statistics (*N* = 203), the difference of misinterpretation percentage of *p*-value between those who got their highest degree overseas (*N* = 52, 73%) and their counterparts who got their highest degree in China mainland (*N* = 196, 86%) was not significant: χ2 (1) = 3.726, *p* = .053, BF10 = 2.073; and the difference of interpretation of CIs is not significant, χ2 (1) = 0.023, *p* = .879, BF10 = 0.231. For those who major in other subjects (*N* = 1276), difference between China mainland (*N* = 1080, 91% for *p*-value and 95% for CIs) and overseas (*N* = 151, 87% for *p*-value and 90% for CIs) for misinterpretation of *p*-value and CIs were not significant, for *p*-value: χ2 (1) = 1.676, *p* = .195, BF10 = 0.171, and for CIs, χ2 (1) = 3.495, *p* = .062, BF10 = 0.362 (Detailed information see **Fig. S1 c & d**)

**Supplementary analysis and result 4: Difference of confidence between correct and wrong answers**

We also used *t* – test to compare the confidence level between correct and wrong answers (See **Table S1** for detailed effect sizes and Bayesian Factors). Although some of the t – test showed there was a significant difference of confidence level between correct answers and wrong answers, it should be noted that all the confidence levels (even wrong answers) were high (all over 3.5 in 5, see Confidence level for wrong answers in **Table S1**) and the effect sizes were too small to interpret.

**Table S1. Difference of confidence between correct and wrong answers.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Question | Error rate | Confidence level(for correct answers) | Confidence level (for wrong answers) | *t* | Cohen’s *d* | BF10 |
| *p*-value A | 53% | 4.07 | 3.88 | 3.34\*\*\* | 0.17 | 14.48 |
| *p*-value B | 48% | 3.97 | 3.79 | 3.09\*\* | 0.16 | 6.48 |
| *p*-value C | 62% | 3.86 | 3.83 | 0.48 | 0.03 | 0.07 |
| *p*-value D | 49% | 3.84 | 3.87 | -0.52 | -0.03 | 0.07 |
| CI A | 57% | 4.13 | 3.93 | 3.60\*\*\* | 0.19 | 34.75 |
| CI B | 54% | 4.01 | 3.95 | 0.95 | 0.05 | 0.09 |
| CI C | 48% | 3.86 | 3.80 | 1.15 | 0.06 | 0.11 |
| CI D | 57% | 3.81 | 3.75 | 1.11 | 0.06 | 0.11 |

*Note:* *p*\*\*\*<.001, *p*\*\*<.005, *p*\*<.05

**Supplementary analysis and result 4: Difference between respondents who had published papers and those did not**

 Among our sample, 534 respondents (36%) had published papers in peer-reviewed journals and the rest didn’t. However, we didn’t find evidence that these two groups are different. For interpretation of *p*-value, 88% respondents had published papers had at least one wrong answer and 90% respondents didn’t published papers had at least one wrong answer, χ2 (1)=0.754, *p*=.385, BF10 = 0.089, and for interpretation of CIs, 93% respondents had published papers had at least one wrong answer and 92% respondents didn’t published papers had at least one wrong answer, χ2 (1)=0.209, *p*=.648, BF10 = 0.039.

For those participants who have published papers, most of them chose to report *p*-value only (*N* = 254, 48%), and 160 participants (30%) reported both *P*-value and CIs in their publication.

**Table S2. Error rate of respondents who have published paper and those not**

|  |  |
| --- | --- |
| Item | Error rate |
| Have published papers | Not published papers |
| *p*-value A | 52% | 54% |
| *p*-value B | 50% | 47% |
| *p*-value C | 62% | 63% |
| *p*-value D | 52% | 47% |
| CI A | 58% | 57% |
| CI B | 56% | 53% |
| CI C | 48% | 47% |
| CI D | 57% | 57% |

**Supplementary analysis and result 5: Results without the highest error rate item (statement C for *p*-value)**

 Among our sample, 924 respondents (62%, the highest error rate) had the wrong answer for the statement C of *p*-value. We analyzed the data without this item again and found almost the same results except lower total error rate. Without this item, 79% respondents had at least 1 error when interpreting *p*-value. For *p*-value without item C, 81% respondents who got their highest degree in China mainland (*N* = 1231) had at least one wrong answer and 71% respondents who got their highest degree overseas (*N* = 248) had at least one wrong answer, χ2(1) = 13.56.38, *p* < .001, BF10 = 62.446.

**Supplementary material: Why these items are incorrect? (Haller & Krauss, 2002; Hoekstra, Morey, Rouder, & Wagenmakers, 2014)**

***P* – value:**

The correct interpretation of a significant *p*-value (α = 0.05) should be:

*The probability of the available (or of even less likely) data, given that the null hypothesis is true, is less than 5%.*

A: You have absolutely disproved (proved) the null hypothesis.

 Item A are easily classified as being FALSE because significance tests can never prove (or disprove) hypotheses. Significance tests provide probabilistic information and can, therefore, at best be used to corroborate theories.

B: You have found the probability of the null hypothesis (alternative hypothesis) being true.

 It is generally impossible to assign a probability to any hypothesis by applying significance tests: One can neither assign it a probability of 100% nor any other probability (α or *p*-value). Therefore, item B should be classified as FALSE. Making statements about probabilities of hypotheses is only possible in the alternative approach of Bayesian statistics.

C: You know, if you decide to (not to) reject the null hypothesis, the probability that you are making the wrong decision.

 The same as item B, it is impossible to assign a probability to any hypothesis by applying significance tests. Item C may look similar to the definition of an error of Type I (i.e., the probability of rejecting the H0 although it is in fact true), but having actually rejected the H0, this decision would be wrong, if and only if the H0 were true. Thus, the probability in item C (“...that you are making the wrong decision”) is *p*(H0), and this probability cannot be derived with NHST.

D: You have a reliable experimental finding in the sense that if, hypothetically, the experiment was repeated a great number of times, you would obtain a significant result on 99% (21%) of occasions.

 Item D reflects the so-called “replication fallacy”. In Neyman and Pearson’s paradigm,

one could interpret α = .01 in a frequentistic framework as relative frequency of rejections of H0 if H0 is true. The example however gives no evidence of the H0 being true. “In the minds of many, 1 - *p* erroneously turned into the relative frequency of rejections of H0, that is, into the probability that significant results could be replicated” (Gigerenzer, 1993a).

**Confidence interval:**

The correct interpretation of a 95 % confidence interval should be:

*If we were to repeat the experiment over and over, then 95 % of the time the confidence intervals contain the true mean.*

A: There is a 95% probability that the true mean lies between 0.1 and 0.4 (-0.1 and 0.4).

B: If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 to 0.4 (-0.1 to 0.4).

D: The null hypothesis is that there is no difference between the mean of experimental group and control group. if you decide to (not to) reject the null hypothesis, the probability that you are making the wrong decision is 5%.

 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.

Similar to item B and C in *p*-value section, item A and D of CI assign probabilities to parameters or hypotheses, something that is not allowed within the frequentist framework. Item B mention the boundaries of the CI (i.e., 0.1 to 0.4, or -0.1 to 0.4), whereas, a CI can be used to evaluate only the *procedure* and not a specific interval.

C: If the null hypothesis is that there is no difference between the mean of experimental group and control group, the experiment has disproved (proved) the null hypothesis.

 Similar to item A in *p*-value section, significance tests can never prove (or disprove) hypotheses. Significance tests provide probabilistic information and can, therefore, at best be used to corroborate theories.