Revisiting the “Belief in the law of small numbers”:   
Conceptual replication and extensions Registered Report   
of problems reviewed in Tversky and Kahneman (1971) [Stage 1]

Cheuk Kiu Hong  
ORCID: 0009-0002-6543-8524  
Department of Psychology, University of Hong Kong, Hong Kong SAR  
cheukkiu@connect.hku.hk / jefferyhong00@gmail.com

^Gilad Feldman  
ORCID: 0000-0003-2812-6599  
Department of Psychology, University of Hong Kong, Hong Kong SAR  
gfeldman@hku.hk / giladfel@gmail.com

^Corresponding author

## Author bios:

Cheuk Kiu Hong was a guided thesis student at the University of Hong Kong during the academic year 2022-3.

Gilad Feldman is an assistant professor with the University of Hong Kong psychology department. His research focuses on judgment and decision-making.

## Declaration of Conflict of Interest:

The author(s) declared no potential conflicts of interests with respect to the authorship and/orpublication of this article.

## Financial disclosure/funding:

The project is supported by the University of Hong Kong Teaching Development Grant.

## Authorship declaration:

Cheuk Kiu Hong conducted the replication as part of his thesis in psychology.

Gilad Feldman guided the project, supervised each step in the project, ran data collection, conducted the pre-registration, and edited the manuscript for submission.

## Corresponding author

Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; [gfeldman@hku.hk](mailto:gfeldman@hku.hk); 0000-0003-2812-6599

## Rights:

CC BY or equivalent license is applied to the AAM arising from this submission. ([clarification](https://bit.ly/rrs-primer))

## Important links and information

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

## Contributor Roles Taxonomy

|  |  |  |
| --- | --- | --- |
| **Role** | **Cheuk Kiu Hong** | **Gilad Feldman** |
| Conceptualization | X | X |
| Pre-registration | X | X |
| Data curation |  | X |
| Formal analysis | X |  |
| Funding acquisition |  | X |
| Investigation | X |  |
| Pre-registration peer review / verification |  | X |
| Data analysis peer review / verification |  | X |
| Methodology | X |  |
| Project administration |  | X |
| Resources |  | X |
| Software | X |  |
| Supervision |  | X |
| Validation |  | X |
| Visualization | X |  |
| Writing-original draft | X |  |
| Writing-review and editing |  | X |

# Abstract

[IMPORTANT:

Abstract, method, and results were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection. For the purpose of the simulation, we wrote things in past tense, but no pre-registration or data collection took place yet.]

The belief in the law of small numbers is the phenomenon that people tend to think small random samples are highly representative of the population they are drawn from. In a Registered Report with an American online Amazon Mechanical Turk sample using CloudResearch (*N* = 1000), we conducted a conceptual replication and extension of seven problems reported by Tversky and Kahneman (1971) examining laypersons’ intuitions. [...]

[The following findings are concluded from simulated random noise and will be updated after data collection.]

We found …

Extending the replication, we found …

We conclude …

Materials, data, and code are available on: <https://osf.io/mns7j/> .

*Keywords:* belief in the law of small numbers, bias, judgment and decision making, registered report, replication, sample size, gambler’s fallacy

# PCIRR Stage 1 Snapshot (revised)

**Provisional title**.

Revisiting the “Belief in the law of small numbers”: Conceptual replication and extensions Registered Report of problems reviewed in Tversky and Kahneman (1971) [Stage 1]

**Authors and affiliations.**

Cheuk Kiu (Jeffery) HONG (cheukkiu@connect.hku.hk), Gilad FELDMAN (giladfel@gmail.com); Department of Psychology, University of Hong Kong

**‎Research question(s) and/or theory.**

We aim to replicate and extend the empirical demonstrations in the review paper regarding people’s intuitions on the “belief in law of small numbers” reviewed in Tversky and Kahneman (1971).

Phenomenon: People hold erroneous intuitions about the laws of chance, both laypersons and scholars and statisticians well-trained in statistics and science. The demonstrations were mostly surrounding the belief in the law of small numbers, that a sample randomly drawn from the population is highly representative of that population, even when the sample size is very small. They had additional demonstrations such as a related belief that a random process is self-correcting, and discussed implications for science and the scientific process.

**Hypotheses/Phenomenon**

We will base our investigation on the demonstrations reviewed by Tversky and Kahneman (1971) with adjustments and extensions and will therefore examine whether we are able to demonstrate the same phenomenon 50 years later using a large diverse layperson sample.

**Study design and methods.**

The target article reviewed several empirical demonstrations of the phenomenon. We aim for a reproduction and replication of 7 of those demonstrations with an adjusted combined design, merging seven reviewed problems into a single unified data collection, presented in random order. We will translate each of the problems to laypersons minimizing jargon and statistical terms (such as z-value, t-value, and statistical significance). Following the original paper, we will compare participants’ answers to the correct answer and document deviations and compare our findings to that of the target.

We will add several measures aimed to clarify the effect and a manipulation of sample size (control: x from target article, mid: 10x, large: 100x) to examine causality.

**Key analyses that will test the hypotheses and/or answer the research question(s).**

The target only used descriptives, which we will try to supplement with statistical tests, such as one-sample proportions. We will use one-way ANOVA with contrasts against the control condition for the sample size manipulation.

**Conclusions that will be drawn given different results.**

We will aim to evaluate the replicability of our findings against the original’s using the Lebel et al. (2019) paradigm. Given the lack of statistical tests, we will compare our adapted statistical tests, and similarity in descriptives, focusing mostly on directionality and whether there was a signal.

**Key references.**

Tversky and Kahneman (1971).<https://doi.org/10.1037/h0031322>

LeBel et al. (2019).<https://doi.org/10.15626/MP.2018.843>

[Notes about changes from previous snapshot and initial plan:

We originally were hoping to run this for both scholar samples and laypersons samples. However, in our other meta projects, it seems that scholars are becoming extremely weary of emails from meta-studies asking them about their scholarly activities and presenting them with questions that may seem to question their understanding of methods and statistics. We were concerned with slow data collection times, very low response rates that are hard to interpret and adjust for, final underpowered small samples, and - alas - growing community annoyance with us and meta-researchers, especially as early-career researchers.

We believe that the attempt to translate scholarly stats questions to laypersons and gauge laypersons’ intuitions about interpreting academic articles, with a new extension manipulating sample size, is already a massive undertaking and complex enough. We there focused the current project on laypersons. We share the adjustments we made to the target’s scholars version, and present it as a mid-way towards the needed translation for laypersons’. We therefore also reframed the project replication classification from a direct replication with extensions to a conceptual replication.

We hope we or others will follow-up on this project with a direct replication of the scholars’ version.]

# PCIRR-Study Design Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Question | Hypothesis | Sampling plan | Analysis plan | Rationale sensitivity of test | Interpretation given different outcomes | Theory that could be shown wrong by the outcomes |
| Do laypeople  ignore sample size in assessing scientific evidence? (perceive population and samples do have the same statistical properties) | Laypersons perceive findings from empirical studies as representative of a phenomenon in the population regardless of the studies’ sample size. | We aimed to recruit 1100 participants aiming for 1000 post exclusions. | One sample t-test and one-way ANOVA with post-hoc analyses. | Aiming for h = 0.20 (far weaker than effects reported in target) multiplied by 3 given 3 conditions, with alpha of 5% and power of 95%, one tail accounting for possible exclusions. This is on par and higher than typical replications in PCIRR. | We examine the replicability of Tversky and Kahneman (1971) examining signal and directionality of effect compared to original. | Belief in the law of small numbers  (contrasted against “Belief in the law of large numbers”, see intro) |
| Given a certain parameter of the population and a randomly drawn sample from that population with cases deviating from the population mean in one direction: laypersons still perceive that the sample’s average to follow that of the population’s average. |
| If a study reports a phenomenon using any sample, then laypersons tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population. | Chi-square test, one sample t-test, and one-way ANOVA with post-hoc analyses. |
| In a study with a sample drawn from the population, laypersons overestimate the importance of power analysis in the likelihood of finding an effect. | One sample t-test and one-way ANOVA with post-hoc analyses. |
| If a study reports a surprising phenomenon using any sample, then laypersons tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population.  Laypersons tend to underestimate the sample size required to confirm the surprising exploratory finding.. | Chi-square test, one sample t-test, and one-way ANOVA with post-hoc analyses. |
| Laypersons do not differentiate between exploratory studies that produce surprising results and confirmatory studies that seek to replicate those surprising results.  Laypersons ignore sample size when comparing exploratory studies and replication studies and consider them to be equally representative of the population regardless of sample size. | Chi-square test, one sample t-test, and one-way ANOVA with post-hoc analyses. |
| Laypersons overestimate the likelihood of confirming exploratory correlational findings with a replication with similar or smaller sample size. |

# Revisiting the “Belief in the law of small numbers”: Conceptual replication and extensions Registered Report of problems reviewed in Tversky and Kahneman (1971) [Stage 1]

## Background

Research by Tversky and Kahneman (1971) demonstrated that people easily mistake the statistical characteristics of a small sample to be highly similar to that of the population, or in other words, that people have a tendency to ignore sample size when evaluating probabilities.

We report a conceptual replication and extension Registered Report, as defined by the criteria of LeBel et al. (2018), of the problems reviewed in Tversky and Kahneman (1971) on the belief in the law of small numbers with the following goals. Our first goal was to conduct an independent well-powered replication of the belief in small numbers phenomenon in the general population. Our second goal was to build on the target’s design and add extensions aiming to: 1) improve on the target’s methods, 2) clarify implicit assumptions in the target’s scenarios, and 3) directly examine the (lack of) causal impact of sample size on representativeness by manipulating the sample size in the reviewed problems.

We begin by introducing the literature on the belief in the law of small numbers and the chosen article for replication - Tversky and Kahneman (1971). We introduce the target’s hypotheses and findings, review our adjustments and extensions, and outline our replication and extension design.

## Belief in the law of small numbers

Tversky and Kahneman (1971) described the “belief in the law of small numbers” as the tendency to perceive small samples randomly drawn from a population as carrying the same statistical characteristics as that of the population. Or, in other words, that a small random sample is highly representative of the population it is drawn from, leading to ignoring sample size in evaluating scientific evidence.

In their seminal article Tversky and Kahneman (1971) provided several straightforward examples for the “belief in the law of small numbers” phenomenon. One of those is about perceptions of IQ distribution. With a population averaging an IQ of 100, when a person from a randomly selected sample from the population is revealed to have an IQ of 150, people still tend to perceive the average IQ of that random sample to be 100, ignoring the provided information regarding that one person which if taken into account would mean an adjustment of the sample’s average IQ to 101. This is likely due to the expectation that the score of any individual is equally likely to be above or below the given average for the population. This is related to the belief that randomness is self-correcting, such that for any person with an IQ of 150 there should be a matching person with an IQ of 50 to balance out the deviation from the population. However, Tversky and Kahneman reiterated that this belief is false given that the characteristics of a population are not reflected in any randomly drawn small samples.

The phenomenon has been the basis for many well-known related biases such as the “Gambler’s fallacy” (Bar-Hillel & Wagenaar, 1991), the belief that the probability of a random event is affected by previous instances of that event. One classic example is with coin flips, in which a fair coin toss with an overall 50% chance of heads or tails leads to the belief that if one landed tails (heads) then the next coin toss is more likely to be heads (tails). The belief is grounded in the expectation that a deviation from the mean towards one end is going to be “offset” by an equal deviation in the opposite direction.

## Choice of article for replication: Tversky and Kahneman (1971)

We chose the article by Tversky and Kahneman (1971) based on several factors: impact, potential for improvements and insights, and relevance for the ongoing science reform.

The article has had a major impact on scholarly research in the area of judgment and decision making, behavioral economics, social psychology, and cognitive psychology. At the time of writing, there were 4585 Google Scholar citations of the article and many important follow-up theoretical and empirical articles.

Tversky and Kahneman (1971)'s work has important practical implications. Given that the phenomenon was originally demonstrated on a sample of scholars and statisticians, who regularly conduct academic research, it inevitably holds important insights for the scientific endeavor and how scholars conduct and evaluate science (e.g., Braver et al., 2014). Bishop et al. (2021) is one of the many papers discussing these implications with follow-up attempts to mitigate this inaccurate belief in the scientific literature with interventions. It also has direct bearing on the issue of employing small samples in psychology and throughout the sciences with insufficient power to detect effects. The growth of meta-science employing the scientific method to the study of science showed concerning trends in size of samples and power employed in psychological science prior to 2015 (Fraley et al., 2022; Fraley & Vazire, 2014; Sassenberg & Ditrich, 2019). One of the goals in the science reform movement has been to improve scientific process has been to directly tackle this issue aiming to correct perceptions about power with a push calling scholars to conduct power and sensitivity analyses which seems to have resulted in larger samples, atleast in psychology (Nosek et al., 2022).

In a fascinating turn of events regarding science credibility and the phenomenon of the “belief in the law of small numbers”, Kahneman, one of the coauthors of the target article, has publicly admitted to falling prey to the very effect he warned the world against. In his mega best-seller book “Thinking fast and slow” (Kahneman, 2011), Kahneman included a chapter on “The Marvels of Priming” enthusiastically reviewing some of the literature on what has later been coined “social priming”, concluding that “We now know that the effects of priming can reach into every corner of our lives”, that “the effects of the primes are robust”, and “the results are not made up, nor are they statistical flukes.” (in chapter “Primes That Guide Us”). A year after the book came out, in 2012, Kahneman wrote an email - later made public - to the lead “social priming” researchers indicating that “count me in as a general believer” calling these researchers to address “questions [that] have been raised about the robustness of priming results”, and “people have now attached a question mark to the field” and therefore that “it is your responsibility to remove it” (Yong, 2012). In 2017, five years later, Kahneman went on a blog summarizing the irreplicability of the findings in his book’s social priming chapter, and wrote a memorable comment regarding his own “belief in law of small numbers” fallacy (Kahneman, 2017):

“What the blog gets absolutely right is that I placed too much faith in underpowered studies. As pointed out in the blog, and earlier by Andrew Gelman, there is a special irony in my mistake because the first paper that Amos Tversky and I published was about the belief in the “law of small numbers,” which allows researchers to trust the results of underpowered studies with unreasonably small samples. We also cited Overall (1969) for showing “that the prevalence of studies deficient in statistical power is not only wasteful but actually pernicious: it results in a large proportion of invalid rejections of the null hypothesis among published results.” Our article was written in 1969 and published in 1971, but I failed to internalize its message. [...]

I knew, of course, that the results of priming studies were based on small samples, that the effect sizes were perhaps implausibly large, and that no single study was conclusive on its own. What impressed me was the unanimity and coherence of the results reported by many laboratories. I concluded that priming effects are easy for skilled experimenters to induce, and that they are robust. However, I now understand that my reasoning was flawed and that I should have known better. Unanimity of underpowered studies provides compelling evidence for the existence of a severe file-drawer problem (and/or p-hacking). The argument is inescapable: Studies that are underpowered for the detection of plausible effects must occasionally return non-significant results even when the research hypothesis is true – the absence of these results is evidence that something is amiss in the published record. [...]

I have changed my views about the size of behavioral priming effects – they cannot be as large and as robust as my chapter suggested.

The lesson I have learned, however, is that authors who review a field should be wary of using memorable results of underpowered studies as evidence for their claims.”

This is a painful and remarkable case-study regarding the “belief in the law of small numbers”, which we took to heart, and yet thought that Kahneman did not go far enough with his conclusions and failed to fully internalize its message regarding his own research. If we revisit his mentioned seminal paper, we find very little details on what has been done, what samples were employed, and in the samples and findings described the samples were very small, and, as far as we know, with no independent well-powered (pre-registered direct) replications. To truly internalize the message regarding Kahneman’s journey with the bias that he was among the first to communicate to the world could have been to also revisit - or call for mass replications of - his own work, even or especially on the “belief in the law of small numbers”. The more we looked into the article, the more questions we had regarding both theory, the argued phenomenon, and the methods, from the need to clarify questions regarding the phenomenon and its robustness, through expected effect size and testing a null hypothesis (e.g., expecting no differences in interpretation for samples of different sizes), and all the way to questions regarding causality (or lack of). There is much need to revisit and clarify these classic findings, especially given Kahneman’s own realization that knowing about the effect does not help mitigate it and the simple fact that his own study on the phenomenon is likely in need of similar remedies to those he has called others to implement in their research.

This holds especially relevant given controversies in the literature in a now classic debate with a different group of scholars maintaining that people actually hold a “belief in the empirical law of large numbers” (Sedlmeier & Gigerenzer, 1997; see discussion below in our extension section), arguing for the need for reframing these effects, identifying possible moderating factors such as study design (single/between versus within) and presentation format. At the very least, before approaching these debates using tools like adversarial collaborations, it would be best to revisit and ensure that findings hold, that they are robust, and that they mean what we concluded them to mean. We previously conducted a replication and extension of and adversarial collaboration of these two groups by Kahneman with Hertwig coordinated by Mellers (Mellers et al., 2001), and found mixed and confusing results in replicating their joint findings (Chandrashekar et al., 2021), showing the importance of revisiting these classics on both sides of the debate to ensure both demonstrations and their resolutions are on solid grounds.

We also thought the article to be especially relevant given recent worrying trends suggesting people often misinterpret evidence provided by a single study with small samples in a specific context, overestimating its generalizability to the larger population (Zhan & Savani, 2023), which recently has become especially crucial in the context of the COVID-19 pandemic (e.g., IJzerman et al., 2020).

We therefore embarked on a well-powered conceptual replication and extension Registered Report of Tversky and Kahneman (1971) aiming to revisit the classic phenomenon with a close translation of the problems reviewed to the context of laypersons and to test the phenomenon in the general population with an extension examining causation by manipulating the described sample sizes.

## Replication closeness evaluation

We provided details on the classification of the replications using the criteria by LeBel et al., (2018) criteria in Table 1 (see section “replication closeness evaluation” in the supplementary materials about the classification). We were strongly grounded in the target article’s claims and empirical demonstration, yet made major changes to allow for a test of its generalizability to laypersons. We summarized the changes we made in Table 3, and decided to classify this as a “far” conceptual replication based on the criteria by LeBel et al. (2018) given our many adjustments to the stimuli from expert language to targeting laypersons’, and the shift in target population.

Another system of categorizing replications was by Nosek and Errington (2020), which puts more emphasis on the match between claims of the replication and the target rather than the match between the replication and the target’s procedures and methods. To be classified as a replication, outcomes from the study that are consistent with the prior claim increase the confidence in the claim, and outcomes that are inconsistent with the prior claim decrease the confidence in the claim. The target article by Tversky and Kahneman (1971) was somewhat vague about the claims, as for example they sometimes referred to scholars and used methods that require some background with statistics and methods, and sometimes the claims seem to be broad and to also make an argument about a wider effect that also holds for the general population.

We see value in both approaches to replications. We use the paradigm by LeBel et al. (2018) to document our deviations regarding process and methods, and at the same time are building on the Nosek and Errington (2020) paradigm in our aim of testing the generalizability of the claims made in the target article to a broader population.

Table 1

*Classification of the replication, based on LeBel et al. (2018)*

|  |  |  |
| --- | --- | --- |
| **Design facet** | **Replication** | **Details of deviation** |
| Effect/hypothesis | Same |  |
| IV construct | Same |  |
| DV construct | Same |  |
| IV operationalization | Same |  |
| DV operationalization | Same |  |
| Population (e.g., age) | Similar with extension | The sample in the target article were mostly people or scholars with statistical knowledge. We targeted laypersons without statistical knowledge. |
| IV stimuli | Adjustments with extension | We added two conditions manipulating the sample size in the questions. |
| DV stimuli | Adjustments with extension | Most of the stimuli were modified or rephrased based on the original questions to make them clearer or to adjust those for our laypersons target sample. |
| Procedural details | Similar, singular unified data collection | Procedures for each or the problems were the same, as far as we can tell from the few details provided.  On a study level, we combined the seven problems into a single data collection using the same participants, presented in random order. |
| Physical settings | Different | The original questionnaire was distributed at the meetings of Mathematical Psychology Group and of the American Psychology Association, and the replication was collected online. |
| Contextual variables | Different | The target article did not report participants’ demographic information. Different time, different settings, different population. |
| Replication classification | “Far” conceptual replication |  |

## Original hypotheses and findings in target article

We aimed to replicate seven of the problems reviewed in Tversky and Kahneman (1971). We marked those as Problems Q1, Q2, Q3, Q4, Q5, Q6, and Q8, according to the order they were presented in the target article.

In Table 2 we summarized the hypotheses of the specific demonstrations by Tversky and Kahneman (1971) and our reframing of the specific example to a broader generalized hypothesis. In our extension explained below, we also manipulated the sample size described in these problems, and in Table 2 we also added a column that reframes the “belief in the law of small numbers” to a causal test null hypothesis for no differences between the conditions with different sample size.

Table 2

*Tversky and Kahneman (1971): Summary of problems*

| Problem | Demonstrations by  Tversky and Kahneman (1971) | Generalized hypothesis (in our replication: laypersons) | Law of small numbers  [Sample size manipulation null hypotheses] |
| --- | --- | --- | --- |
| Q1:  ”Replication success” | Perceiving a study with a sample of 20 and a replication of that study with a sample of 10 to show similar results. | Laypersons perceive findings from empirical studies as representative of a phenomenon in the population regardless of the studies’ sample size. | Laypersons perceive a study with a sample of X, 10 times X, and 100 times X, and replications of those studies with a sample of X/2, 10X/2, and 100X/2 show similar results. Regardless of what X is. |
| Q2: “IQ” | Given a population with an average IQ of 100, and a randomly drawn sample of 50 from that population with the first individual having an IQ of 150: People still perceive the overall average IQ of the drawn sample to be 100.  [Associated assumption: A single deviation in one direction is offset by similar deviation in the opposite direction.] | Given a certain parameter of the population and a randomly drawn sample from that population with cases deviating from the population mean in one direction: laypersons still perceive that the sample’s average to follow that of the population’s average.  [Associated assumption: Deviations in one direction are offset by similar overall deviations in the opposite direction.] | Given a population with an average IQ of 100, and a randomly drawn sample of X from that population with the first X/50 cases having an average IQ of 150: Laypersons still perceive the overall average IQ of the drawn sample to be 100. Regardless of what X is.  [Associated assumption: Average deviations of a subgroup in one direction are offset by similar overall deviations in the opposite direction.] |
| Q3:  “Infants” | If a study reports that 4 out of 5 infants preferred Toy A over Toy B, then people tend to perceive that as representative of the general population and therefore that most infants in the general population prefer Toy A to Toy B. | If a study reports a phenomenon using any sample, then laypersons tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population. | If a study reports that 80% of X infants preferred Toy A over Toy B, then laypersons tend to perceive that as representative of the general population and therefore expect that about 80% infants in the general population will prefer Toy A to Toy B. Regardless of what X is. |
| Q4:  “Population correlation” | If population effect is estimated to be r = .35, and power analysis indicated a required sample of 79. In a study with a sample of 79 meeting a power analysis, people tend to expect the same effect size as when the sample is larger.  [See clarification note below] | In a study with a sample drawn from the population, laypersons overestimate the importance of power analysis in the likelihood of finding an effect. | If population effect is estimated to be r=Y, and power analysis indicated required sample of X. In a study with a sample size larger than X, laypersons tend to expect the same likelihood for the effect size Y. Regardless of what X is. |
| Q5:  “Exploratory analyses” | If a study with a sample of 40 reports support for an exploratory surprising finding, then people tend to perceive that as representative of the general population and therefore for that exploratory finding to hold true for the population.  People tend to underestimate the sample size required to confirm the exploratory finding (by about half - 20). | If a study reports a surprising phenomenon using any sample, then laypersons tend to perceive their findings to be representative of the general population and therefore expect that the finding generally holds true for the general population.  Laypersons tend to underestimate the sample size required to confirm the surprising exploratory finding. | If a study with a sample of X reports support for an exploratory surprising finding, then laypersons tend to perceive that as representative of the general population and therefore for that exploratory finding to hold true for the population. Regardless of what X is.  Laypersons tend to underestimate the sample size required to confirm the exploratory finding. Regardless of what X is. |
| Q6:  “Failed replication” | If a study with a sample of 40 reports support for an exploratory surprising finding, and a confirmatory replication with a sample of 20 fails to find support for that finding, people tend to perceive both studies as equally likely and representative of the population requiring finding a reason for the differences. | Laypersons do not differentiate between exploratory studies that produce surprising results and confirmatory studies that seek to replicate those surprising results.  Laypersons ignore sample size when comparing exploratory studies and replication studies and consider them to be equally representative of the population regardless of sample size. | If a study with a sample of X reports support for an exploratory surprising finding, and a confirmatory replication with a sample of X/2 fails to find support for that finding, laypersons tend to perceive both studies as equally likely and representative of the population requiring finding a reason for the differences. Regardless of what X is. |
| Q8: “20 variables, 190 correlations” | If an exploratory study with a sample of 100 found support for 27 correlations, then people overestimate the likelihood of finding support for these associations in a replication with a sample of 40 (expecting 66%+ successful replication). | Laypersons overestimate the likelihood of confirming exploratory correlational findings with a replication with similar or smaller sample size. | If an exploratory study with a sample of X found support for Y correlations, then laypersons overestimate the likelihood of finding support for these associations in a replication with a sample of 40% of X. Regardless of what X is. |

*Note*. Q6 is a follow-up question to Q5. We skipped Question 7 as it is covered by the scenario described in Q5 and Q6.  
Regarding Q4: “Population correlation” - The target article probably referred to a “critical significance value” table, and not a power analysis. As far as we know, these are no longer in use, and most scholars today use power analysis. We understand that this changes what the study is about, yet decided to run this adjusted to a power analysis to try and determine whether people indeed understand power analyses impact on probabilities of finding.   
The demonstrations by Tversky and Kahneman targeted a mix of populations, yet given our focus on laypersons we framed the generalized hypothesis to specifically mention laypersons to be clear about the target population. Laypersons refers to people with limited knowledge of statistics who do not use statistics in their everyday life.

## Extension: Causal effect of sample size

The core message of the target article is that people neglect to take sample size into consideration when evaluating sample representativeness of the population. The demonstrations in the target article were mostly using descriptives with vignettes describing single sample size scenarios. These were powerful demonstrations, and yet they fell short of truly demonstrating that people do not take sample size into account, because there was no manipulation of sample size in these scenarios.

As a first step, we summarized our translation of the hypotheses in the specific demonstrations of the target article in Table 2. For each of those scenarios manipulating sample size there are competing hypotheses regarding how the manipulation would impact people’s evaluations: the “law of small numbers” phenomenon seems to suggest support for the null hypothesis - that sample size manipulations would not impact the interpretation of findings.

The alternative hypothesis is that people do take sample size into account. One theoretical account that does not assume a null-hypothesis is the belief in the “law of large numbers” which predicts that the larger the sample, the more likely people are to perceive it as representative of the population (Sedlmeier & Gigerenzer, 1997). Previous research demonstrated that people can intuitively infer that the larger the sample size the more likely it is to resemble the characteristics of the population, with youngsters as early as the age of 11 showing indications of having this intuition (Piaget & Inhelder, 1951/1975) and initial similar demonstrations in adults (Peterson & Beach, 1967).

To try and resolve the two seemingly opposite findings, Sedlmeier and Gigerenzer (1997) suggested that it might be dependent on the type of question, such as a distinction between frequency distribution (“a distribution of values from one sample”, p. 36) and sampling distribution (“distribution of means from independent samples of fixed size, drawn from the same population”). In so doing, it seems that Sedlmeier and Gigerenzer (1997) accepted the premise that the results of the target article hold and generalize, and that their resolution relies on the statistical presentation of the problem. Yet, it is possible that the phenomenon is dependent on the experimental design, and that a direct manipulation of the sample size in these scenarios would allow us a stricter empirical causal test to contrast “belief in the law of small numbers” suggesting null differences against the possibility of “belief in the law of large numbers” suggesting an adaptation to evaluations based on sample size. In the case that people do adjust their evaluations according to sample size, a manipulation would allow us an initial estimation of the direction and the extent to which people make such adaptation. The implementation of that manipulation across the different scenarios would allow us to examine whether such adjustments differ based on the question at hand. Together, these may suggest a more nuanced view of the debate regarding the seemingly contradictory findings.

## Pre-registration and open-science

We provided all materials, data, and code on: <https://osf.io/mns7j/> . This project received Peer Community in Registered Report Stage 1 in-principle acceptance ((ENTER LINK AFTER IPA); (ENTER LINK AFTER IPA)) after which we created a frozen pre-registration version of the entire Stage 1 packet (ENTER LINK AFTER IPA) and proceeded to data collection. All measures, manipulations, exclusions conducted for this investigation are reported, and data collection was completed before analyses.

# Method

## Power and sensitivity analyses

We first calculated effect sizes of the findings reported in the target article, then conducted an a priori power analysis (power = 0.95, alpha = 0.05). The reported effects in the target article were large (Q1: *h* = 0.90; Q5: *h* = 0.87; Q6/D: *h* = 0.32; Q2/3/4/8: no information) and therefore the minimum sample size required based on smallest effect in Q6/D was 125. We supplemented this analysis with setting the SESOI to *h* = 0.20, considered a weak effect, requiring a sample of 325 participants, about 2.5 times the required sample based on the target’s effects, on par with adjustment recommendations by Simonsohn (2015) (even though meant for other designs).

We tripled the sample size given the extension of adding two conditions to the independent variable (=972), added margins to compensate for the exploratory interaction, and for any potential data exclusions. As a result, we concluded we would aim for a sample size of 1100 participants, 366 per each of the four conditions, likely 1000+ overall and 333+ per condition after exclusions.‎

A sensitivity analysis indicated that a sample of 1000 (after exclusions) would allow the detection of *f* = 0.12 for a three conditions between-subject design ANOVA in our experimental design (95% power, alpha = 5%). Also, the sample would be sufficiently powered to detect contrasts of *d* = 0.33 (200 per condition, 95% power, alpha = 5%, one-tail), which are effects much weaker than those reported in the target article and correspond to typical medium effects in social psychology research (Xiao et al., 2023).

## Participants

[Note: To demonstrate the results after data collection we simulated a dataset of 1000 participants using Qualtrics and reported our analyses below based on that dataset. Results will later be updated to the final sample with the real data.]

We recruited a total of 1000 US American participants on Amazon Mechanical Turk using CloudResearch (Litman et al., 2017) (*Mage* = M.MM, *SD* = SD,SD; X males, X females, X others/rather not disclose). Based on our extensive experience running similar judgment and decision-making replications on MTurk, we will employ the following CloudResearch options: Duplicate IP Block to ensure high-quality data collection. Duplicate Geocode Block, Suspicious Geocode Block, Verify Worker Country Location, Enhanced Privacy, CloudResearch Approved Participants and Block Low-Quality Participants. We will also employ the Qualtrics fraud and spam prevention measures: reCAPTCHA, prevent multiple submissions, prevent ballotstuffing, bot detection, security scan monitor and relevantID, etc.

[The assignment pay is based on the federal wage of 7.25USD/hour, per minute, so for example 5-8 minutes survey would be paid 1 USD per participant. We first pretested survey duration with 30 participants to make sure our time run estimate was accurate and adjusted pay as needed, the data of the 30 participants was not analyzed other than to assess survey completion duration and needed pay adjustments. For those pretest participants, if survey duration was longer than expected, they were paid a bonus as pay adjustment. The pretest participants' responses were included in the final analysis.]

## Design and procedure

[For review: The Qualtrics survey .QSF file and an exported DOCX file are provided on the OSF folder. A preview link of the Qualtrics survey is provided on:

<https://hku.au1.qualtrics.com/jfe/preview/previewId/b0b9be99-3191-4ae4-9c7d-41212b7f2f92/SV_bDWVv5m9EXpqgxo?Q_CHL=preview&Q_SurveyVersionID=current> ]

We summarized the stimuli and experimental design in Table 3. We provided additional information about the target’s original stimuli and our adjustments in the supplementary materials’ “Adjustments to the target article’s original questions” subsection.

We reconstructed the target’s stimuli and adjusted it to an online Qualtrics survey based on the information provided in the article. Participants indicated their consent, with four questions confirming their eligibility, understanding, and agreement with study terms, which they must answer with a “yes” and required responses in order to proceed to the study. Three of the four questions had the options order being rotated (yes, no, not sure), thereby also serving as attention checks by ensuring that participants carefully read the question options. Participants who failed to indicate their consent and answer the rotated consent questions were asked to return the task.

Participants were randomly assigned to one of three conditions manipulating the sample size described in the scenarios (see extension section below). Within each condition, participants answered a total of seven questions, Problems Q1, Q2, Q3, Q4, Q5, Q6, and Q8 from the target article (by order of mention in the article). The display of the problems within each condition was randomized (Qualtrics’ “evenly presented”), except for Problem Q5 and Problem Q6 which were grouped together since Q6 was a follow-up scenario to Q5.

We note we did not include Q7 as it covered a similar theme to the questions we added to Q5 and Q6 and referred to “value of t” requiring statistical knowhow that we found difficult to translate to laypersons. Specifically, the target’s Q7 was as follows:

“An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, *t* = 2.46. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of *t* with the same number of subjects. The direction was the same in both sets of data.

You are reviewing the literature. What is the highest value of *t* in the second set of data that you would describe as a failure to replicate?”

Problem Q7 therefore aimed to get at one’s perception of the statistical threshold for “t” that needs to be met to conclude a failure to replicate. In both Problem Q5 and Problem Q6 we added questions on top of those presented in the target which aim at ideas related to those from Problem Q7, asking about confidence in the findings, required sample size, etc.

We chose to combine all questions into a unified single data collection given that they were different enough and tapped different aspects of the phenomenon. The method of combining several scenarios of the same phenomenon to a single data collected was previously tested successfully in many of the replications and extensions conducted by our team (e.g., Yeung & Feldman, 2022; Zhu & Feldman, 2023), and is especially powerful in addressing concerns about the target sample (e.g., naivety and attentiveness) when some studies replicate successfully whereas others do not, as well as in allowing for drawing inferences about links between the different scenarios and examining the consistency in participants’ responding to similar decision-making paradigms. At the end, participants answered a number of funneling questions and provided their demographic information, and then debriefed.

Table 3

*Tversky and Kahneman (1971) problems: Original version and our adaptation for laypersons*

|  | Original version | | Laypersons version | | |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Q1: ”Replication success”** | "Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory (*z* = 2.23, *p*< .05, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?" | | You read a media report about an experiment that was run on [**20/200/2000**] people, and the report indicates that the result shows support for their theory.  The same scientists ran the same experiment again (a replication) with [**10/100/1000**] people of the same population.  (Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)  [Likelihood scale: 0-100] | | |  |
| **Q2: “IQ”** | The mean IQ of the population of eighth graders in a city is known to be 100. You have selected a random sample of 50 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample? | | The average IQ of all eighth graders in a city is reported to be 100. A scientist randomly selected [**50/500/5000**] children from all those eighth graders to investigate their educational achievements. The [**first child/average IQ of the first 10/100 children**] in the group of [**50/500/5000**] children [has an IQ of/is] 150.  [Text input: 1-200] | | |  |
| **Q3: “Infants”** | …suppose he is engaged in studying which of two toys infants will prefer  to play with. Of the first five infants  studied, four have shown a preference for the same toy. | | **DV1: Rejection of null hypothesis (replication)**  Based on this evidence, can the scientists conclude that infants overall have a preference for Toy A over Toy B?  [Multiple choice (yes/no)]  **DV2: Replication likelihood (extension)**  If the scientists were to conduct another study with [**10/100/1000**] infants, what is the likelihood that [**8/80/800**] or more out of the [**10/100/1000**] would prefer Toy A over Toy B?  (0% = Absolutely will NOT happen; 100% = Absolutely will happen)  [Likelihood scale: 0-100]  **DV3: required sample size (extension)**  Suppose that [**8/80/800**] out of the [**10/100/1000**] infants in the scientists’ second study also preferred Toy A over Toy B. If these scientists were to run one final study to conclude once and for all that 4 out of 5 infants in the world population prefer Toy A over Toy B, what is the minimum number of infants that the scientists would need to include in that final study? (Try your best to estimate.)  [Numerical text input min: 1, 0 decimal] | | |  |
| **Q4:**  **“Population correlation”** | Imagine a psychologist who studies the correlation between need for Achievement and grades. When deciding on sample size, he may reason as follows: "What correlation do I expect? r = .35. What *N* do I need to make  the result significant? (Looks at table.) *N* = 33. Fine, that's my sample." The only flaw in this reasoning is that our psychologist has forgotten about sampling variation, possibly because he believes that any sample must be highly representative of its population. However, if his guess about the correlation in the population is correct, the correlation in the sample is about as likely to lie below or above .35. Hence, the likelihood of obtaining a significant result (i.e., the power of the test) for N = 33 is about .50.  [*Note: The target article probably referred to a “critical significance value” table, and not a power analysis. As far as we know, these are no longer in use, and most people today use power analysis. We understand that this changes what the study is about, yet decided to run this adjusted to a power analysis to try and determine whether people indeed understand power analyses impact on probabilities of finding*] | | Scientists who study two personality traits (Trait A and Trait B) expect there to be a positive association (relationship) between these two traits in the general population. In other words, scientists expect people with higher ratings on Trait A to also tend to have higher ratings on Trait B. On the possible range of associations from -1 (fully negative association) to 0 (no association) and 1 (fully positive association), the expected association in the general population is 0.35.  You have just read a media report about a new study on the association between Trait A and Trait B. In that study, the scientists conducted an analysis to try and determine how many people need to participate in their study in order to convincingly be able to conclude support for a relationship between Trait A and Trait B, and determined that in order to be able to detect an association of 0.35, they would need to run a study with at least 79 participants.  (Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)  **DV1: Likelihood with estimated sample size (replication)**  If the scientists run their study with 79 participants, what do you think is the likelihood that the scientists will find support for an association of 0.35?  (Reminder: 0.35 is the expected association from the entire population they sampled from)  (0% = Absolutely will NOT happen; 100% = Absolutely will happen)  [Likelihood scale: 0-100]  **DV2: Reasons for not finding support for an effect (extension)**  In your opinion, what are the most plausible reasons why the positive relationship between Trait A and Trait B in their research **might not** be the same as the effect in the population?  [Text input]  Try and estimate: If the scientists ran their study with [**79/790/7900**] participants [(10/100 times larger than in the first study)]…  (0% = Absolutely will NOT happen; 100% = Absolutely will happen)  [Likelihood scale: 0-100]  **DV3: Likelihood of correlation equal population (extension)**  What is the likelihood that in the scientists’ new study with [**79/790/7900**] people, there will be an association of **exactly 0.35** between Trait A and Trait B?  [Likelihood scale: 0-100]  **DV4: Likelihood of correlation larger than population (extension)**  What is the likelihood that in the scientists’ new study with [**79/790/7900**] people, there will be an association **equal to or larger than 0.35** between Trait A and Trait B?  [Likelihood scale: 0-100]  **DV5: Likelihood of correlation larger than population (extension)**  What is the likelihood that in the scientists’ new study with [**79/790/7900**] people, there will be an association of **exactly 0.40** between Trait A and Trait B?  [Likelihood scale: 0-100]  **DV6: Likelihood of correlation smaller than population (extension)**  What is the likelihood that in the scientists’ new study with [**79/790/7900**] people, there will be an association equal to or smaller than 0.35 between Trait A and Trait B?  [Likelihood scale: 0-100]  **DV7: Likelihood of correlation smaller than population (extension)**  What is the likelihood that in the scientists’ new study with [**79/790/7900**] people, there will be an association of exactly 0.30 between Trait A and Trait B?  [Likelihood scale: 0-100] | | |  |
| **Q5:**  **“Exploratory analyses”**  Scenario | Considering the importance of the result, its surprisal value, and the number of analyses that your student has performed—  Would you recommend that he replicate the study before publishing? If you recommend replication, how many animals would you urge him to run ? | | Considering the potential importance of the findings, the surprising results, and the number of factors included in the experiment, try and estimate:  (0% = Absolutely will NOT happen; 100% = Absolutely will happen)  **DV1: Replication likelihood if same person reran (extension)**  What is the likelihood that if the same scientist were to run the same study again on the same number of new people ([**40/400/4000**]) from the same population, that they would obtain the same results?  [Likelihood scale: 0-100]  **if someone else reran (extension)**  What is the likelihood that if someone else were to run the same study again with a new sample of the same size ([**40/400/4000**]), they would obtain the same results?  [Likelihood scale: 0-100]  **DV3: Should the study be rerun (replication)**  Do you think the scientist should try and rerun their study before attempting to publish the results?  [Multiple choice: (yes/no)]  **DV4: Required sample size (replication)** (only applicable when “yes” is chosen for DV3)  If yes - how many people should the study rerun include?  (please take into consideration that funding is limited, running participants is costly, and try and aim for an optimal use of funding and researcher time)  [Text numerical input: min:1, 0 decimal]  **DV5: Confidence in effect (extension)**  If the scientist were to run the same study again on the same number of new people ([**40/400/4000**]) from the same population and found similar results, how confident would you be that their findings represent a real phenomenon that could also be found in the general population?  (0% = Not at all certain; 100% = Absolutely certain)  [Likelihood scale: 0-100] | | |  |
| **Q6:**  **“Failed replication”**  Scenario | 1. should pool the results and publish his conclusion as fact. (0) 2. He should report the results as a tentative finding. (26) 3. He should run another group of [median = 20] animals. (21) 4. He should try to find an explanation for the difference between the two groups. (30) | | **DV1: Recommended course of action (replication)**  What do you think that the scientist should do now?  For each of the options indicate the extent to which you think this is a recommended course of action   1. The scientist should combine the results from the two experiments and publish the results of both studies combined. 2. The scientist should report the results from the two experiments separately and summarize them as tentative findings. 3. The scientist should again rerun the study with a new group of people. 4. The scientist should try and explain the differences between the findings of the two studies. 5. The scientist should ask another scientist to conduct another independent study.   [Scale: -100 = *Strongly recommend against doing*;  100 = *Strongly recommend in favor of doing*]  **DV2: Required sample size (extension)**  Regarding options C and E, if a scientist were to run another third group of new people from the same population, how many participants do you think the scientist should recruit for the new study?  (please take into consideration that funding is limited, running participants is costly, and try and aim for an optimal use of funding and researchers’ time)  [Text input: min:1]  **DV3: Helpfulness of the rerun (extension)**  To what extent did running the second study help with getting closer to the truth and gaining a better understanding of the phenomenon?  [Scale: 0 = *Did not help at all;* 100 = *Helped very much*] | | |  |
| **Q8: “20 variables, 190 correlations”**  Scenario | You have run a correlational study, scoring 20 variables on 100 subjects. Twenty-seven of the 190 correlation coefficients are significant at the .05 level; and 9 of these are significant beyond the .01 level. The mean absolute level of the significant correlations is .31, and the pattern of results is very reasonable on theoretical grounds. How many of the 27 significant correlations would you expect to be significant again, in an exact replication of the study, with *N* = 40 ? | | You read a news story about a study in the field of personality and social psychology that examined the associations (relationships) between 20 different personality traits, and the study was conducted on [**100/1000/10000**] people.  In examining the 20 traits, there are 190 possible associations (association of trait 1 to trait 2,3,4…20, trait 2 to traits 3,4,5…20… up to associations between trait 19 and trait 20).  Of the 190 possible associations, the study found support for 27 associations, and with slightly stronger associations for 9 of those 27 associations.  On the possible range of associations from -1 (fully negative association) to 0 (no association) and 1 (fully positive association), the average of the 27 supported associations was .31. Overall, the scientist thought that the pattern of relationships seemed reasonable and consistent with their theory.  (Clarification: What scientists typically mean by "finding support" is that if they observe a predicted pattern of results in the data, that there is a 5% or lower chance that in reality there is no such true pattern and that this result is due to chance. Scientists then take this data as evidence that increases their confidence in rejecting the idea that there is no such pattern in the population.)  [Text input: min:1 max:27]  [Text input: min:1 max:9]  [Text input: min:1 max:163] | | |  |

*Note*. The current project is focused on the laypersons’ version. Measures are marked as either a replication directly translated for laypersons from the target’s scenarios, or as an extension to indicate that we added this measure to try and further clarify the effect.  
The numbers in parentheses indicate the changes made for the x10 and x100 conditions, with the x10 condition being 10 times the target’s stated sample size, and the x100 condition being 100 times the target’s stated sample size.  
Our adjustments to the original’s targeted at scholars is provided in the supplementary, to help explain the steps taken from the original to our laypersons’ version.

## Exploratory extensions

### Sample size manipulation

Participants were randomly assigned to one of three conditions: control, X10, or X100. Participants in the control condition were presented with questions similar to those in the target article, with our translation to laypersons, using the target’s sample size numbers. We added two conditions: in the x10 condition participants read about samples 10 times the target’s stated sample size, and in the x100 condition participants read about samples 100 times the target’s stated sample size. For example, in Problem Q2 the target article indicated “4 out of 5 infants” to have a preference for Toy A over Toy B. For that problem in the extension conditions, we manipulated them to be 40 out of 50 infants for the X10 condition and 400 out of 500 infants in the X100 condition.

We note that in the scenarios we manipulated all the mentioned sample sizes in the question. For example, in Problem Q1, we varied both the sample size of the described original experiment of 20 (to 200 and 2000), and the sample size of the described replication of 10 (to 100, and 1000). An alternative approach could have been to only vary the described original experiment or to only vary the described replication. We decided to vary both because we wanted to keep the ratio between the original and the replication constant, to be able to examine the overall use of sample size information.

In addition, manipulating sample size may have additional factors that we had not taken into account, as for example reviewer Dr./Prof. Kariyushi Rao suggested the possibility that in Problem Q2 the likelihood of one person having an IQ of 150 in a sample of 50 is higher than the likelihood of an IQ average of 150 in the first 10 people in a sample of 500. We noted the complexity inherent in the manipulation of Problem Q2 yet to keep things simple, we decided not to further complicate the X10 and X100 conditions.

In this extension, we aim to test the competing hypotheses of expected differences between the three conditions. Tversky and Kahneman’s “belief in the law of small numbers” would for the most part predict that people are not sensitive to sample size, and therefore that there should be no detectable differences between the conditions. A competing hypothesis, related to the “belief in the law of large numbers” and mentioned in the introduction, is that there will be differences between the three conditions and that people will update their answers in response to sample size differences.

Our main analyses and empirical focus is on using Null Hypothesis Significance Testing, and this is what we use in order to test for detecting a signal (rejecting the null hypothesis). The conclusions of whether a hypothesis was supported or not will rely on NHST. Yet, in cases in which we fail to reject the null, we will complement all our analyses with Bayesian analyses aimed to try and quantify the evidence in support of the null.

Bayesian analyses are tricky specifically because they incorporate a subjective measure of a prior, which is especially challenging given the competing hypotheses, a replication of an old classic that has not been subjected to many replications, very little in the follow-up literature, and no experience for this research question with our target sample. Therefore, any prior is debatable, and so we implemented a prior of 0.707 that is often used as the default in many Bayesian tools and packages (such as JASP, BayesFactor, and ggstatsplot), generally meant to address ambiguous cases like this. Bayes factor will be reported in our figures, using the R package ggstatsplot, as a supplementary indicator to quantify support for or against the null.

### Statistical background

We added a section asking participants to indicate their level of statistics knowledge (“How would you rate your proficiency in use of statistics?”; 0 = *Not at all proficient in statistics*, 100 = *very proficient in statistics*), use (“How often do you use statistics and statistical inferences in your job?”; 0 = *Not at all*, 100 = *All the time*), and training (“Do you have any training in statistics?”; No / Highschool / College / Professional / Academic).

### Statistics intuitions dependent variables

We summarized the extensions for the statistics intuitions in Table 4.

Table 4

*Extensions mapping and analysis strategy*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Factor | Original problem | Added to problems | Reason for adding | Test |
| Replication success likelihood | Q1 DV1 | Q3 DV2 Q5 DV1-DV2 | Generalizability | One-sample t-test (control)  ANOVA contrasting X, 10X, 100X  Mixed ANOVA contrasting Q5 DV1-2 |
| Estimations of required sample size to find effect (laypersons’ power analysis) | None | Q3 DV3 Q5 DV4  Q6 DV2 | Intuitions regarding required samples sizes to draw inferences | One-sample t-test (control)  ANOVA contrasting X, 10X, 100X |
| Reasons for not finding support for an effect | None | Q4 DV2 | Exploratory. Examining an untested assumption | Qualitative |
| Likelihood of effects | Q4 DV1 | Q4 DV3-DV7 | Likelihood estimations regarding specific associations | Comparing point to range estimations  Prediction: Larger-equal>equal=larger  Prediction: Smaller-equal>equal=smaller  Mixed ANOVA |
| Contribution of replications to confidence in phenomenon | None | Q5 DV5  Q6 DV3 | Whether replications contribute to substantiating a phenomenon | One-sample t-test (control)  ANOVA contrasting X, 10X, 100X |
| Replicability of correlations | Q8 DV1 | Q8 DV2  Q8 DV3 | Examining all reported findings in scenario (supported, strongly supported, and not supported) | ANOVA contrasting X, 10X, 100X |

#### Replication success likelihood

The target article’s Q1 ”Replication success” was aimed at assessing the way people estimate the likelihood of a successful replication. We aimed to further explore those estimations in the other problems and therefore added replication likelihood questions to: 1) Q3: “Infants” (DV2); 2) Q5: “Exploratory analyses” (DV1 and DV2).

The target article’s Problem Q5 focused on whether the original study should be rerun and with what sample size. We aimed to supplement that with examining the perceived likelihood that a replication would show the same results, and whether it would make any difference whether the replication was conducted by the same experimenters or by an independent lab. Therefore, we will conduct a mixed ANOVA (within:same scholar vs. different scholar; between: sample size manipulation) examining laypersons’ evaluations of independent verifications, that results are more likely to be replicated if conducted by the same scientists than by other scholars. This direction is exploratory, and framed as assuming differences, yet we suspect that people might underestimate possible scientist biases and perceive the two as fairly similar (null hypothesis, effect size Cohen’s *d* < 0.2).

#### Estimations of required sample size

We aimed to further explore laypersons’ intuitive power analyses to assess their estimations regarding required sample size to test empirical questions. We added questions to: 1) Q3: “Infants” (DV3), 2) Q5: “Exploratory analyses” (DV4), 3) Q6: “Failed replication” (DV2).

#### Likelihood of effects

The original Problem Q4: “Population correlation” focused on people’s expectations regarding effects similar to that of the population given an estimated required sample size. We wanted to assess not only the likelihood of finding support for an effect but more specific expectations regarding found effects. We included both specific effects, one the same as the population (0.35), one higher than that of the population (0.40), and one lower than the population (0.30), as well as ranges - equal or higher than population (0.35), equal or lower than population (0.35).

This direction is exploratory, yet we suspect that people might tend to overestimate the likelihood of very unlikely exact point estimates, especially the one closest to the population (0.35) and underestimate the likelihood of the much more likely range estimates, especially that with weaker effects than that estimated in the population (equal or lower than 0.35).

## Measures

We provided the exact measures in Table 3, with each measure marked as either a replication directly translated for laypersons from the target’s scenarios, or as an extension to indicate that we added this measure to try and further clarify the effect. In Table 5, we also summarized all predictions, correct/expected answers, effects found in the target article, effects we found in our replication ([To be updated in Stage 2]), and our replication interpretation of whether the findings were supported or not ([To be updated in Stage 2]).

## Evaluation criteria for replication findings

Wherever possible, we aimed to compare the replication effects with the original effects in the target article using the criteria set by LeBel et al. (2019) (see section “Replication evaluation” in the supplementary). Given that the information provided by the target article was sometimes insufficient, and given our many adjustments and extensions, we will revert to indicating a “signal” for whether a prediction was supported or not, and a “direction” indicating whether it was in the expected direction or not.

Table 5

*Control condition: Predictions, correct answers, target article findings, our findings, and interpretation*

[To be updated in Stage 2]

| **Problem** | **Dependent measure** | **Replication/ Extension** | **Prediction** | **Correct answer** | **Target article** | **Effect** | **Interpretation** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Q1:**  **”Replication success”** | Replication likelihood | Replication | People perceive much higher likelihood of success than warranted (ignoring sample size). | 0.48 | Median = 0.85 | [TBD] | [Supported/Not supported] |
| **Q2: “IQ”** | Estimated mean IQ | Replication | People tend to answer this incorrectly (same as in the population = 100 | 101 | Most people answer 100 | [TBD] | [Supported/Not supported] |
| **Q3: “Infants”** | Rejection of null hypothesis | Replication (reframed) | People tend to conclude support and reject the null. | Insufficient to reject null | Many psychologists reject null | [TBD] | [Supported/Not supported] |
| Replication likelihood | Extension | People tend to conclude high likelihood for replication | 0.375 |  | [TBD] | [Supported/Not supported] |
| Required sample size | Extension | People | 95% power = 28 | Exploratory | [TBD] | [Supported/Not supported] |
| **Q4:**  **“Population correlation”** | Likelihood with original sample size | Replication |  |  | They used a sample of 33, claim 0.5 because it is as equally likely to be below and above .35. | [TBD] | [Supported/Not supported] |
| Reasons for not finding an effect | Extension | Exploratory open-ended | N/A | Extension  N/A | Exploratory qualitative. Will not be analyzed. | Exploratory qualitative. Will not be analyzed. |
| DV3-DV7: Various likelihoods | Extensions | Exploratory | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| **Q5:**  **“Exploratory analyses”** | Likelihood if same person rerun | Extension | Exploratory | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| Likelihood if someone else rerun | Extension | Exploratory | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| Should the study be rerun | Replication | People tend to favor replication | Exploratory (surprising) findings should be confirmed. | 66 out of 75 chose to run the replication. | [TBD] | [Supported/Not supported] |
| Required sample size | Replication | People tend to underestimate the required sample size |  | Median is N = 20 | [TBD] | [Supported/Not supported] |
| Confidence in effect | Extension | Exploratory | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| **Q6:**  **“Failed replication”** | Recommended course of action | Replication | Option a: 0/77 = 0%  Option b: 26/77 = 33.77%  Option c: 21/77 = 22.27%  Option d: 30/77 = 38.96%  Option e: extension, not applicable | OptionD “The scientist should try and explain the differences between the findings of the two studies.”  is wrong. | Similar to the original findings, from options a to d, a would have the lowest score, and d should have the highest. Option e is exploratory. | [TBD] | [Supported/Not supported] |
| Required sample size | Extension | Exploratory.  Compared to Q5 “Required sample size”. | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| Helpfulness of the rerun | Extension | According to the law of belief in small numbers, one would expect running a replication with half the sample size should still yield similar results, so it should not provide too much insight into understanding the phenomenon. | Overall, with small samples, running additional successful replications helps increase confidence in the phenomenon. | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| **Q8: “20 variables, 190 correlations”** | Replicability of the 27 supported correlations | Replication | People tend to overestimate the number of rerun supported correlations | T&K: 8 to 10 of the original 27 is “probably a generous estimate” | Median = 18 | [TBD] | [Supported/Not supported] |
| Replicability of the 9 slightly stronger correlations | Extension | Exploratory | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |
| Replicability of the 163 unsupported correlations | Extension | Exploratory | N/A | Extension  N/A | Exploratory: quantitative. | Exploratory: quantitative. |

## Data analysis strategy

### Replication

Eight of the measures in the seven problems are replication dependent variables, taken from the original study with a translation to laypersons aiming to demonstrate the generalizability of the underlying phenomenon in a sample of laypersons. We will therefore compare our findings to that reported in the target’s (for those that reported sufficient details).

The replication tests will be conducted on the control condition. We decided to set the alpha to .01, as .05 is a common threshold used in replications to evaluate a signal in support of the target’s findings, and there are seven problems, with one main replication dependent variable each, therefore slightly higher than the conservative Bonferroni .05/8 suggestion and a round clear alpha target.

### Extensions

We set our alpha threshold to .001 for all extension analysis.

We added extension dependent variables to Problems Q3, Q4, Q5, Q6, and Q7, summarized in Table 4, with up to 7 analyzed variables per each of the problems (when combining the replication with the extension dependent variables). Therefore, .001 meets the strict Bonferroni .01/7 suggestion and a round clear number.

For the extension manipulating the sample size (x versus 10x versus x100), we will conduct a series of one-way ANOVAs with post-hoc contrasts against the control condition for each of the measures. For the ANOVA analyses we will report Holm corrections for multiple analyses and will report both raw and corrected p-values, yet our criteria for signal will use the corrected p-values against the .001 alpha threshold.

### Inclusion criteria, outliers, and winsorizing

In our data analysis we will only include responses from participants who completed the entire questionnaire, have passed the consent checks at the beginning of the questionnaire (participants cannot proceed to the survey without correctly answering those), and rated with “seriousness” >=3 (on a scale of 1 to 5) and English understanding >=4 (on a scale of 1 to 7) in the funneling section at the end of the questionnaire.

We pre-register that in case we fail to find support for the core hypotheses in our replication of the target article, we will then supplement our analyses with rerunning the analyses with outlier winsorizing. Following peer review and a suggestion from reviewer Prof. Romain Espinosa we will employ winsorizing using the R package *datawizard* (Patil et al., 2022) using the method zscore with a cutoff of 3 standard deviations plus or minus the mean. In such a case we will report findings of both before and after winsorizing, and document differences in the findings.

### Robustness checks: Outlier handling

Following feedback from peer review, we will conduct an initial outliers analysis on Problems Q1, Q2 and Q3 (replication DV1) to examine the differences regarding the conclusions from analyzing the full data and from analyzing the data with winsorizing. In case the conclusions for the full sample are in support of the findings and the winsorized sample is not in support of the findings, we will conduct exploratory analyses to examine possible explanations for how the participants whose responses were winsorized differ from the larger sample (attentiveness, demographics, etc.), and will conduct robustness checks with full versus winsorized comparisons for all effects reported in the manuscript. Regardless, our conclusions regarding whether we found support for the effect will depend on the findings using the full sample, yet we will add a section in the general discussion to discuss limitations and implications. Our reasoning for this choice is because we restricted the range of response for all items which makes all responses reasonable, and our experience with the target sample is that it is of high quality responding and attentiveness, and therefore outliers are likely to represent real and thoughtful responses.

# Results

[IMPORTANT:   
Method and results were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection. For the purpose of the simulation, we wrote these sections in past tense, but no pre-registration or data collection took place yet.]

[Please see the Rmarkdown files in the OSF folder for the initial data analysis plan. The results section is provided as a starting point placeholder, and will be updated with further details in Stage 2.]

## Replication

Descriptive statistics of all measures are presented in Table 6 and plotted in Figure 1. Outlier tests of all measures are presented in Table 7. Statistical tests of the hypotheses of the DVs that are categorized as replications are summarized in Table 7.

Table 6

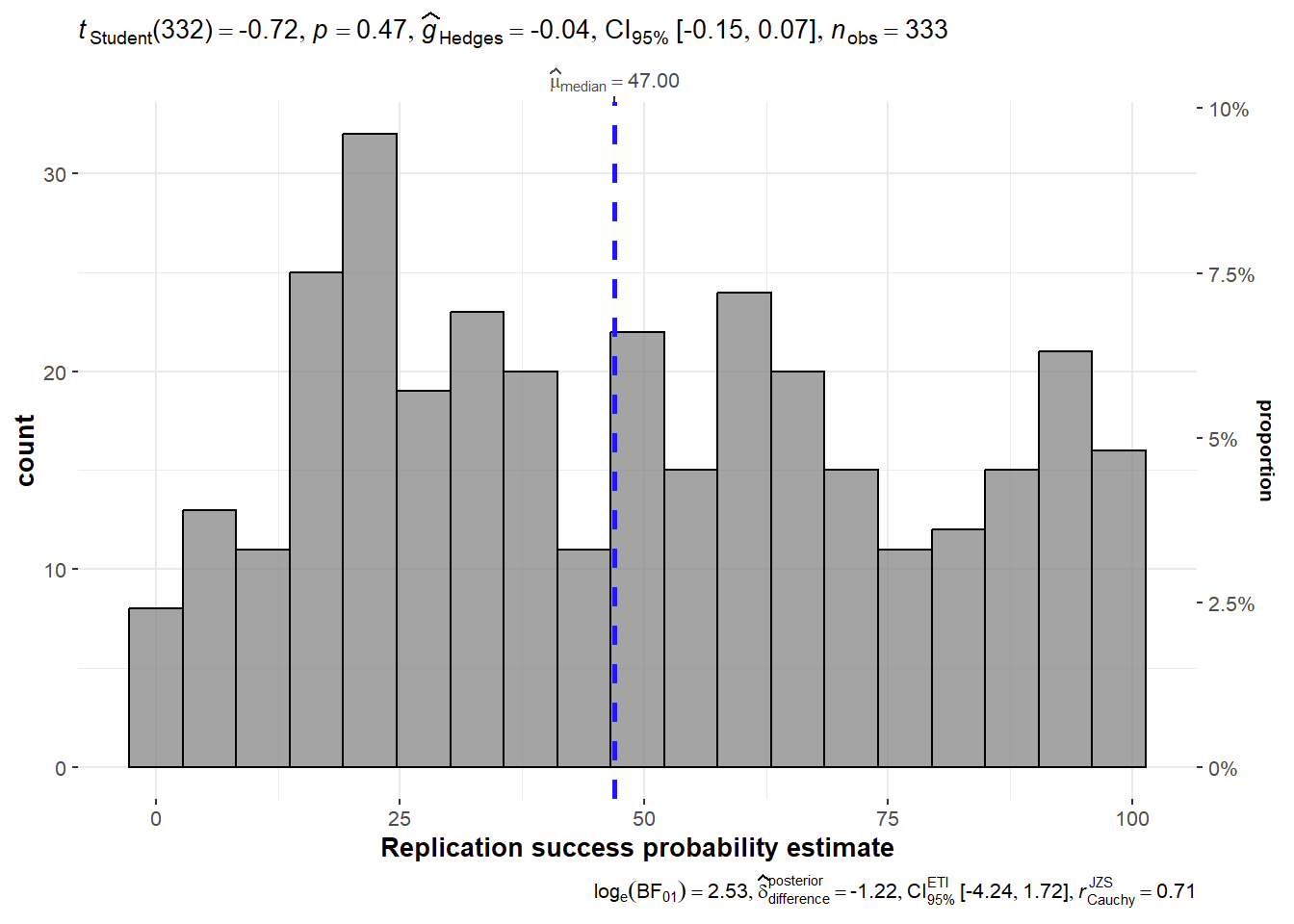
*Descriptive statistics for all conditions*

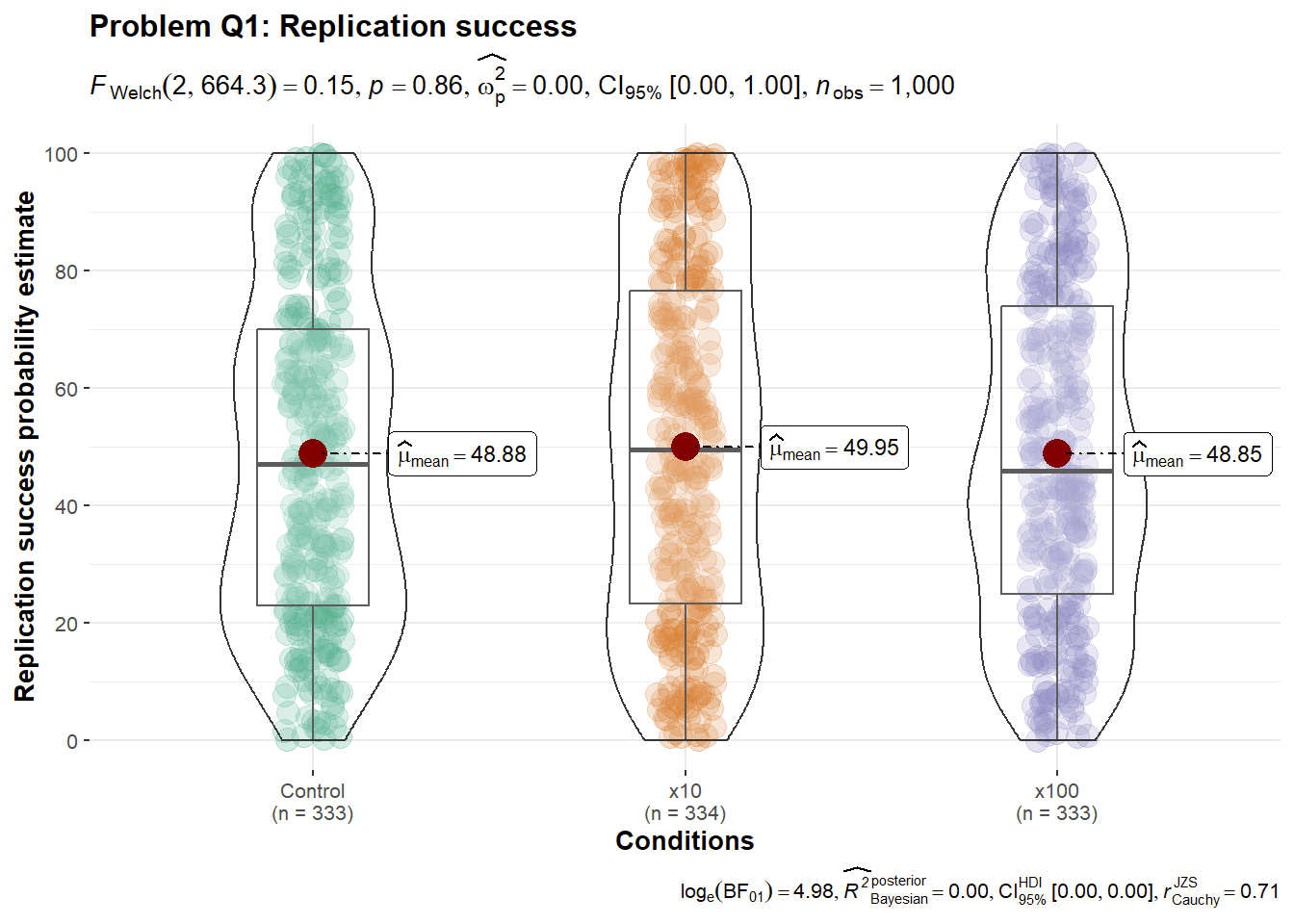
| Problem | DV | Control condition | x10 condition | x100 condition | Overall |
| --- | --- | --- | --- | --- | --- |
| Q1:  ”Replication success” | Replication likelihood | 52.72 [29.39] (165) | 51.80[28.95] (164) | 52.91 [28.45] (164) | 46.23 [29.69] (493) |
| Q2: “IQ” | Estimated mean IQ | 96.19 [56.53] (165) | 100.17 [59.84] (164) | 94.18 [57.40] (164) | 99.17 [55.66] (493) |
| Q3:  “Infants” | Rejection of null hypothesis | 0.42 [0.49] (165) | 0.55 [0.50] (164) | 0.55 [0.50] (164) | 0.51 [0.50] (493) |
|  | Replication likelihood | 48.30 [28.77] (165) | 48.01 [29.32] (164) | 49.90 [29.81] (164) | 48.74 [29.25] (493) |
|  | Required sample size | 50.82 [29.17] (165) | 49.15 [29.82] (164) | 49.75 [30.79] (164) | 49.91 [29.88] (493) |
| Q4:  “Population correlation” | Likelihood with original sample size | 48.07 [30.37] (165) | 51.48 [28.36] (164) | 52.94 [28.05] (164) | 50.82 [28.97] (493) |
|  | Likelihood of replication correlation equal population | 47.83 [29.01] (165) | 54.80 [28.10] (164) | 49.05 [28.41] (164) | 50.55 [28.61] (493) |
|  | Likelihood of replication correlation larger than population | 47.75 [28.23] (165) | 54.80 [28.51] (164) | 48.13 [30.02] (164) | 49.09 [28.92] (493) |
|  | Likelihood of replication correlation larger than population | 51.30 [28.50] (165) | 48.65 [29.32] (164) | 48.65 [28.72] (164) | 50.91 [28.84] (493) |
|  | Likelihood of replication correlation smaller than population | 53.36 [28.71] (165) | 47.97 [28.46] (164) | 49.57 [28.63] (164) | 50.31 [28.63] (493) |
|  | Likelihood of replication correlation smaller than population | 49.77 [28.48] (165) | 47.88 [29.48 (164) | 52.89 [28.38] (164) | 50.18 [28.80] (493) |
| Q5:  “Exploratory analyses” | Likelihood if same person rerun | 47.74 [29.21] (165) | 49.24 [28.73] (164) | 52.37 [28.67] (164) | 49.78 [28.88] (493) |
|  | Likelihood if someone else rerun | 51.28 [27.42] (165) | 52.09 [30.02] (164) | 50.62 [29.50] (164) | 51.33 [28.95] (493) |
|  | Should the study be rerun | 0.53 [0.50] (165) | 0.56 [0.50] (164) | 0.49 [0.50] (164) | 0.53 [0.50] (493) |
|  | Required sample size | 45.65 [27.53] (88) | 46.04 [30.44] (92) | 49.75 [30.01] (80) | 47.05 [29.29] (260) |
|  | Confidence in effect | 48.30 [30.86] (165) | 48.53 [28.02 (164) | 49.90 [30.13] (164) | 48.91 [29.64] (493) |
| Q6:  “Failed replication” | Recommended course of action | A: -2.55 [56.36] (165)  B: -5.43 [57.77] (165)  C: -8.28 [62.54] (165)  D: 4.29 [59.17] (165)  E: 1.79 [62.20] (165) | A: 2.28 [57.18] (164)  B: 5.43 [55.50] (164)  C: -9.07 [56.12] (164)  D: 5.43 [58.04] (164)  E: -7.22 [61.13] (164) | A: -0.39 [54.43] (164)  B: -2.90 [60.77] (164)  C: 3.12 [60.91] (164)  D: -2.68 [56.72] (164)  E: 5.62 [57.69] (164) | A: 0.04 [55.92] (493)  B: -4.35 [57.94] (493)  C: -4.75 [60.06] (493)  D: 2.35 [57.98] (493)  E: 0.07 [60.49] (493) |
|  | Required sample size | 52.08 [27.44] (165) | 53.09 [29.26] (164) | 54.42 [28.64] (164) | 53.19 [28.41] (493) |
|  | Helpfulness of the rerun | 48.81 [30.23] (165) | 52.05 [30.4g5] (164) | 51.16 [28.31] (164) | 50.67 [29.65] (493) |
| Q8: “20 variables, 190 correlations” | Replicability of the 27 supported correlations | 14.07 [8.00] (165) | 12.79 [8.08] (164) | 13.35 [7.72] (164) | 13.40 [7.94] (493) |
|  | Replicability of the 9 slightly stronger correlations | 4.70 [2.77] (165) | 4.84 [2.87] (164) | 4.69 [3.00] (164) | 4.74 [2.87] (493) |
|  | Replicability of the 163 unsupported correlations | 82.61 [48.08] (165) | 86.30 [49.75] (164) | 83.02 [48.57] (164) | 83.98 [48.73] (493) |

*Note*. [MM.MM indicate mean. SD.SD indicate standard deviation. N indicates sample size for that box). Add any needed notes.]

**Figure 1a**

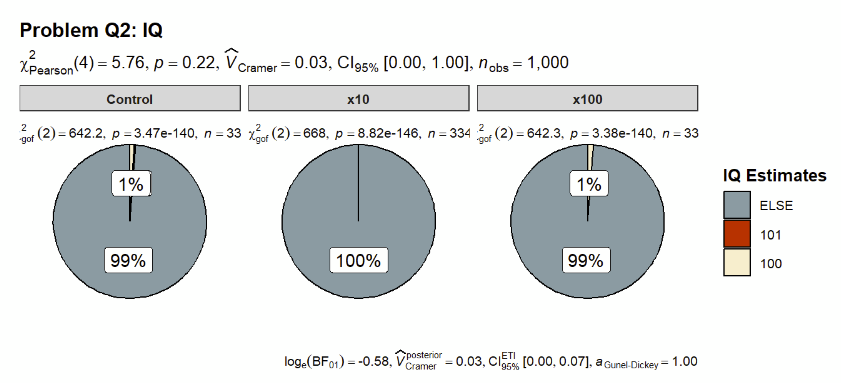
*Problem Q1 ”Replication success”*

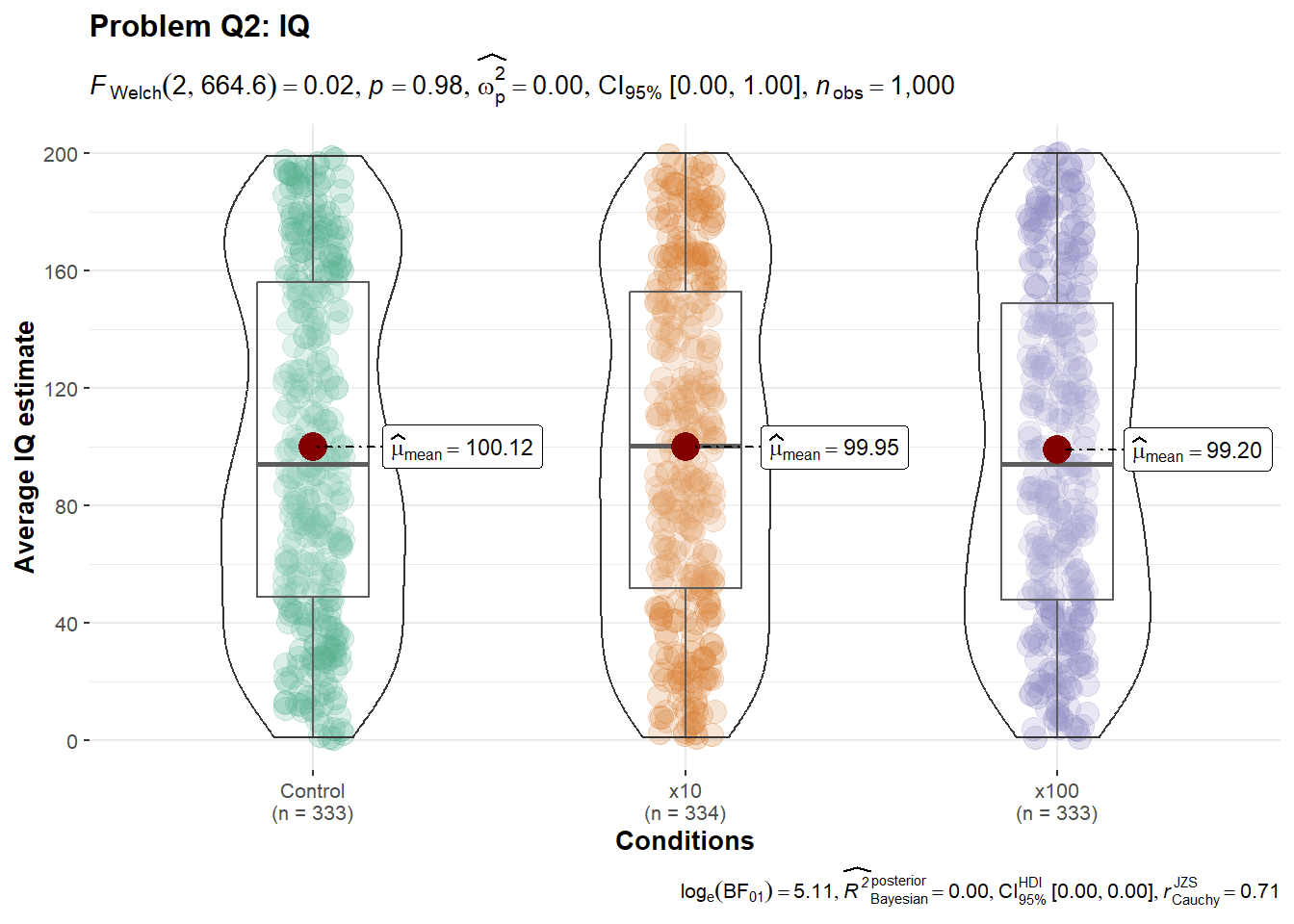




**Figure 1b**

*Problem Q2 “IQ”*

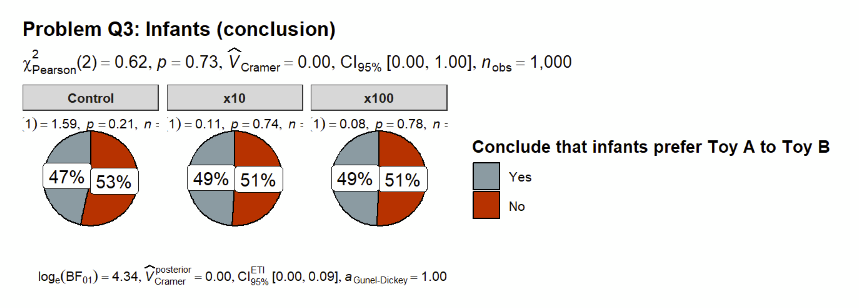


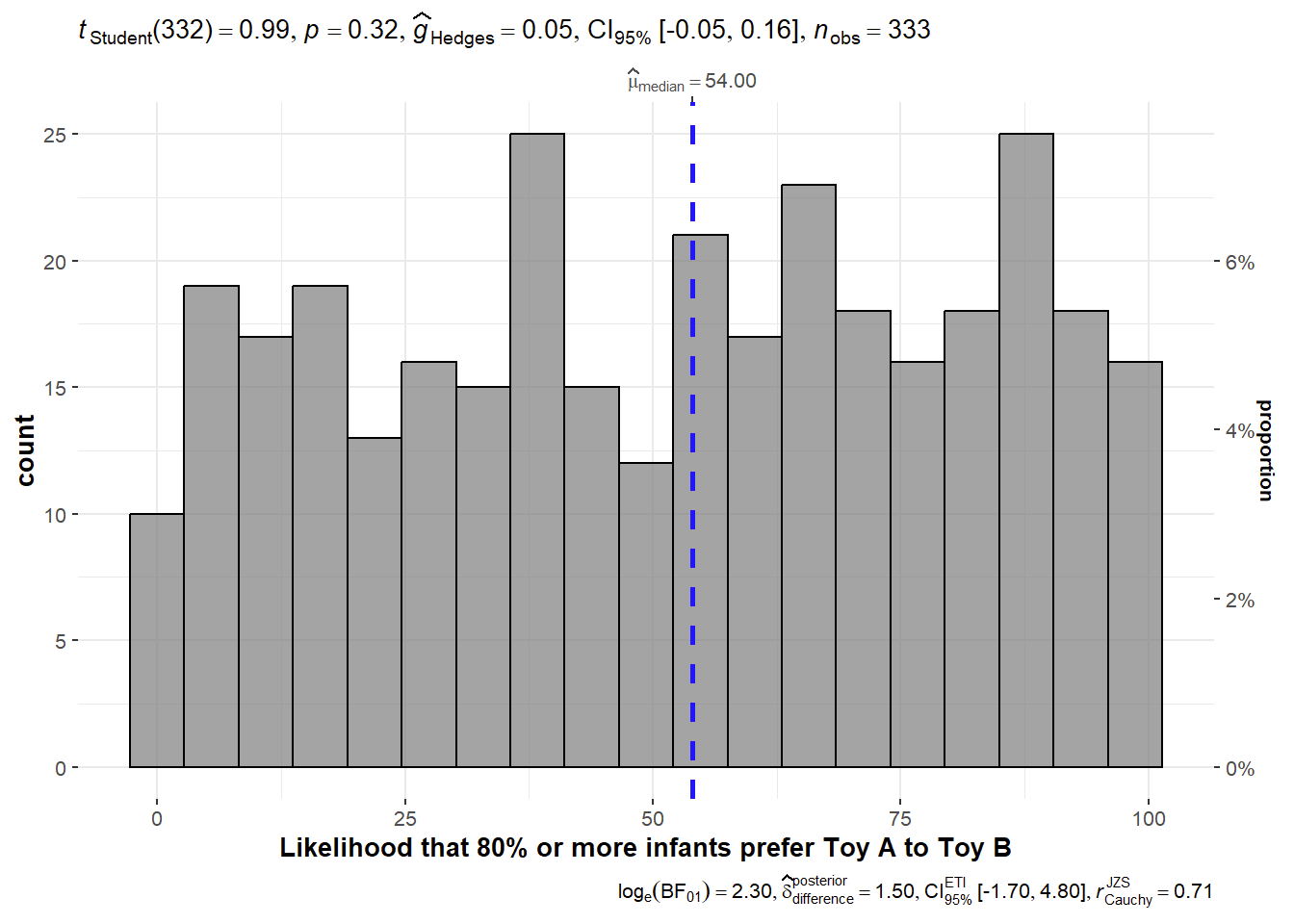


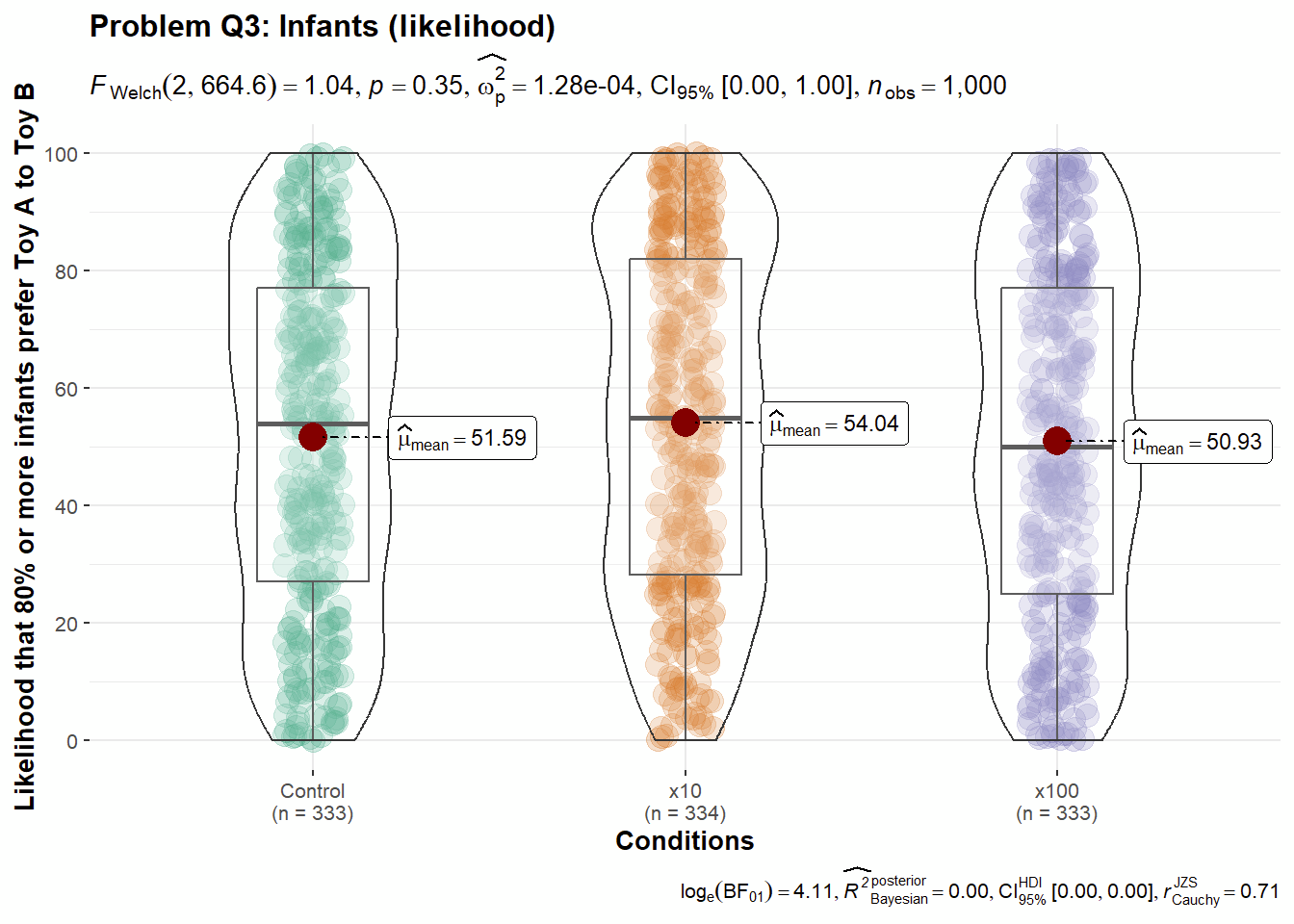
.

**Figure 1c**

*Problem Q3 “Infants”*

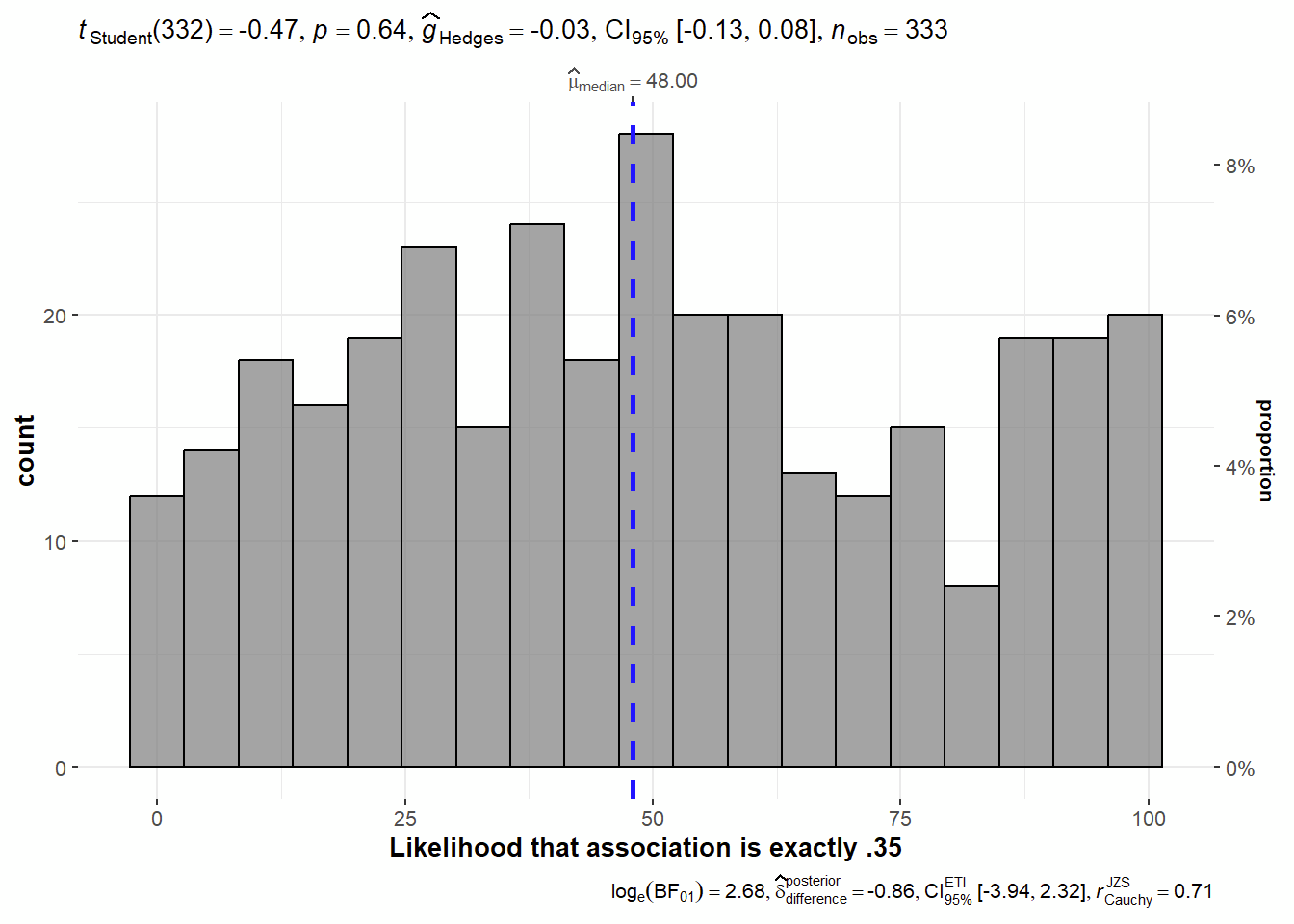


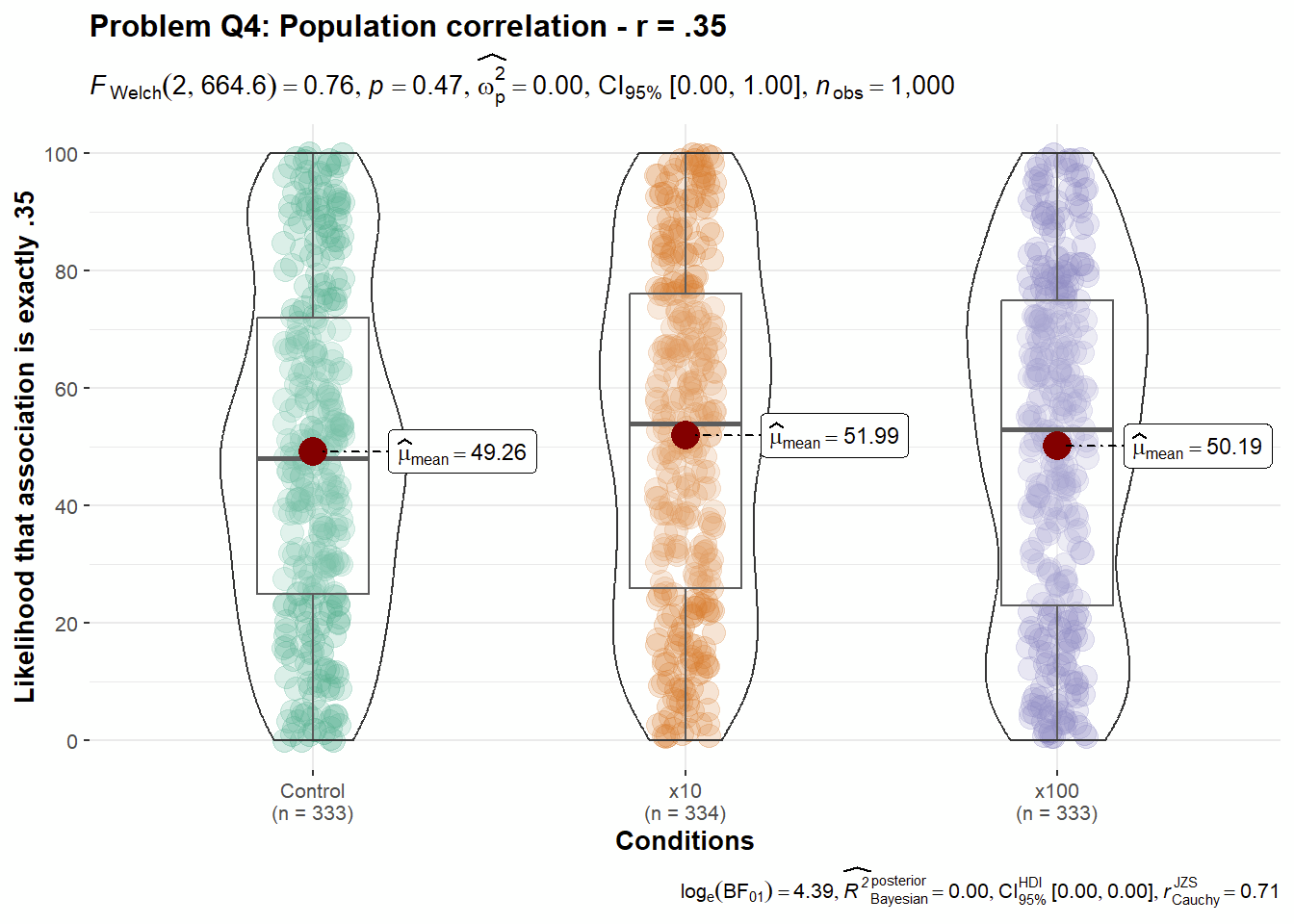




**Figure 1d**

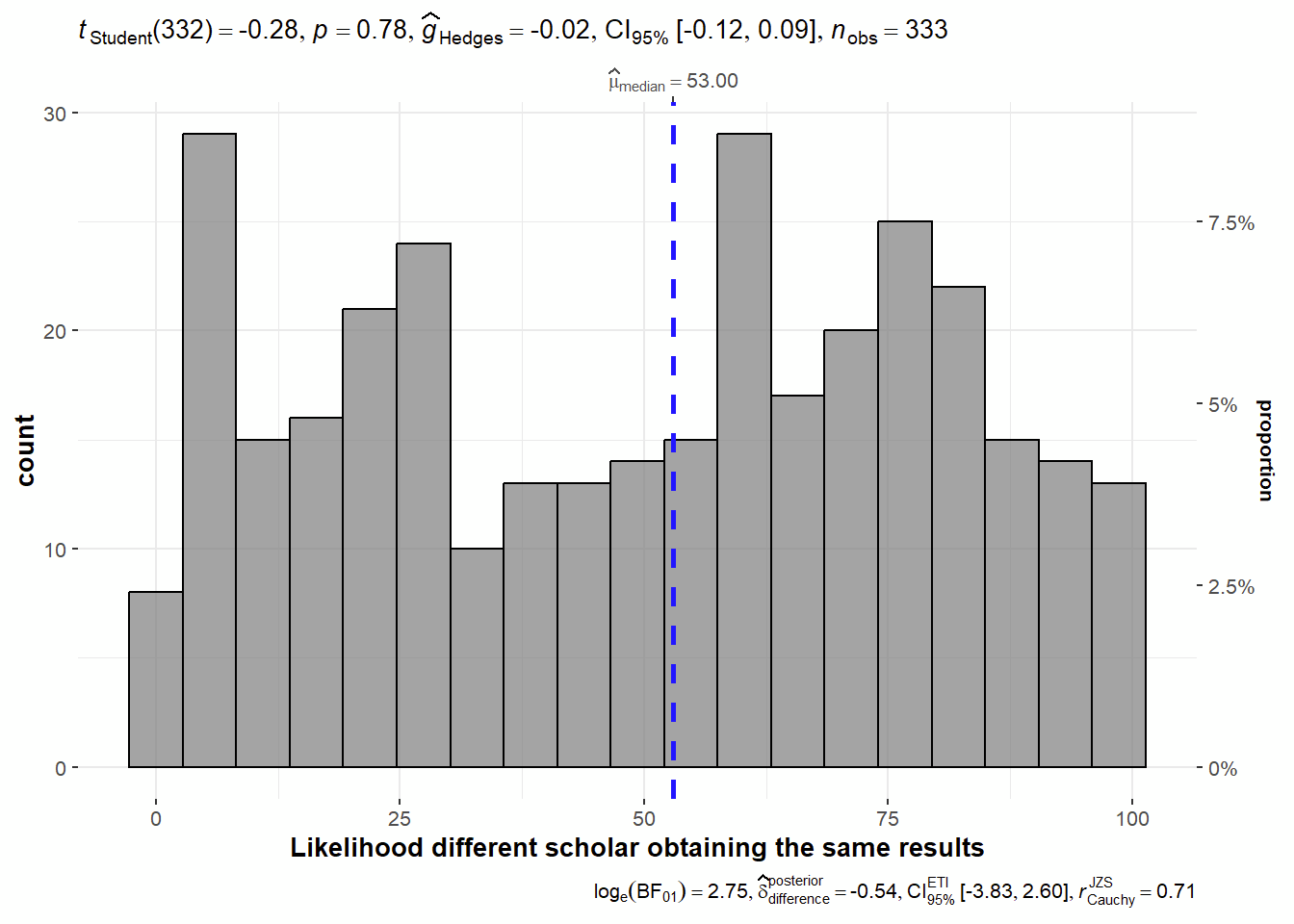
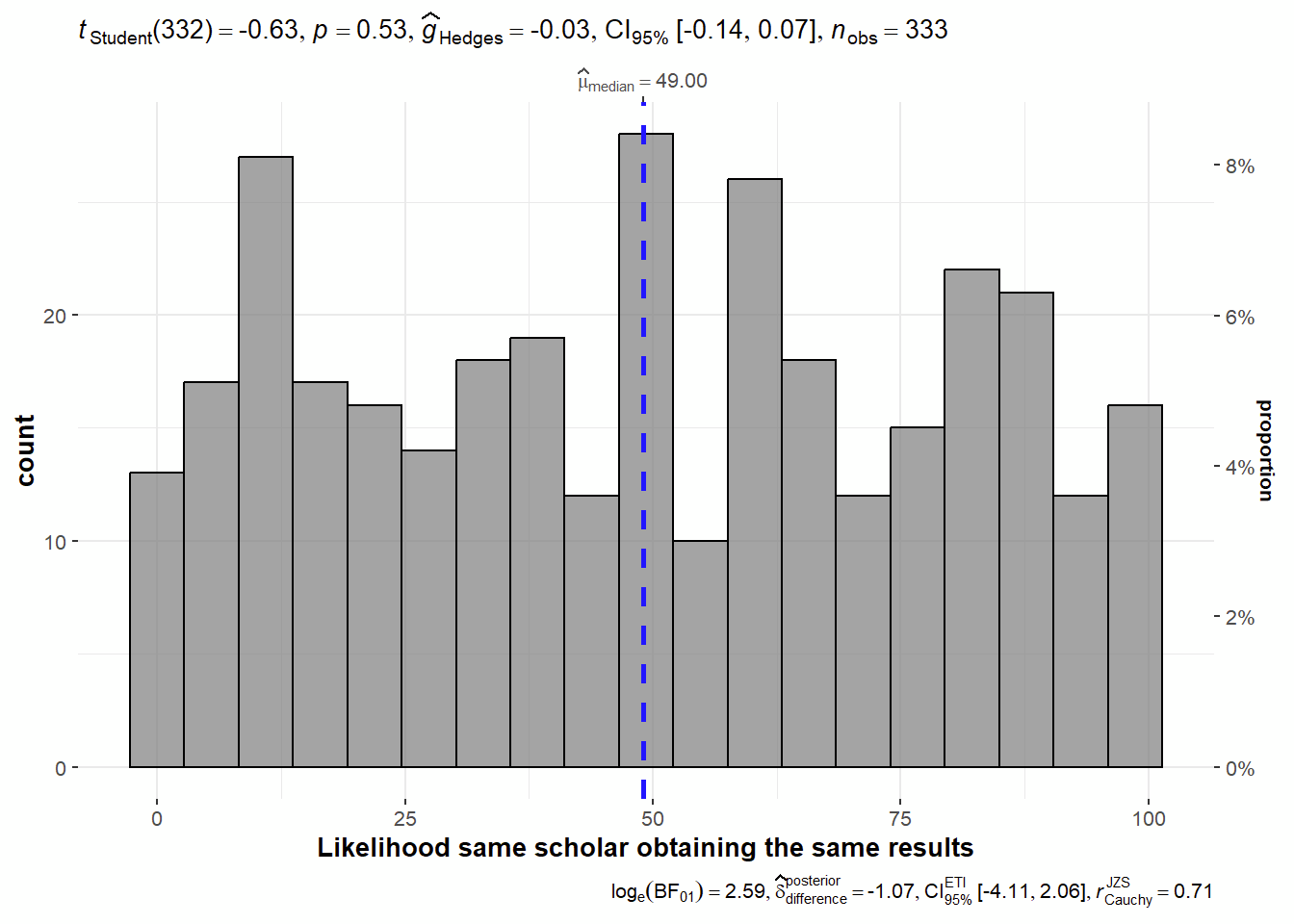
*Problem Q4 “Population correlation”*

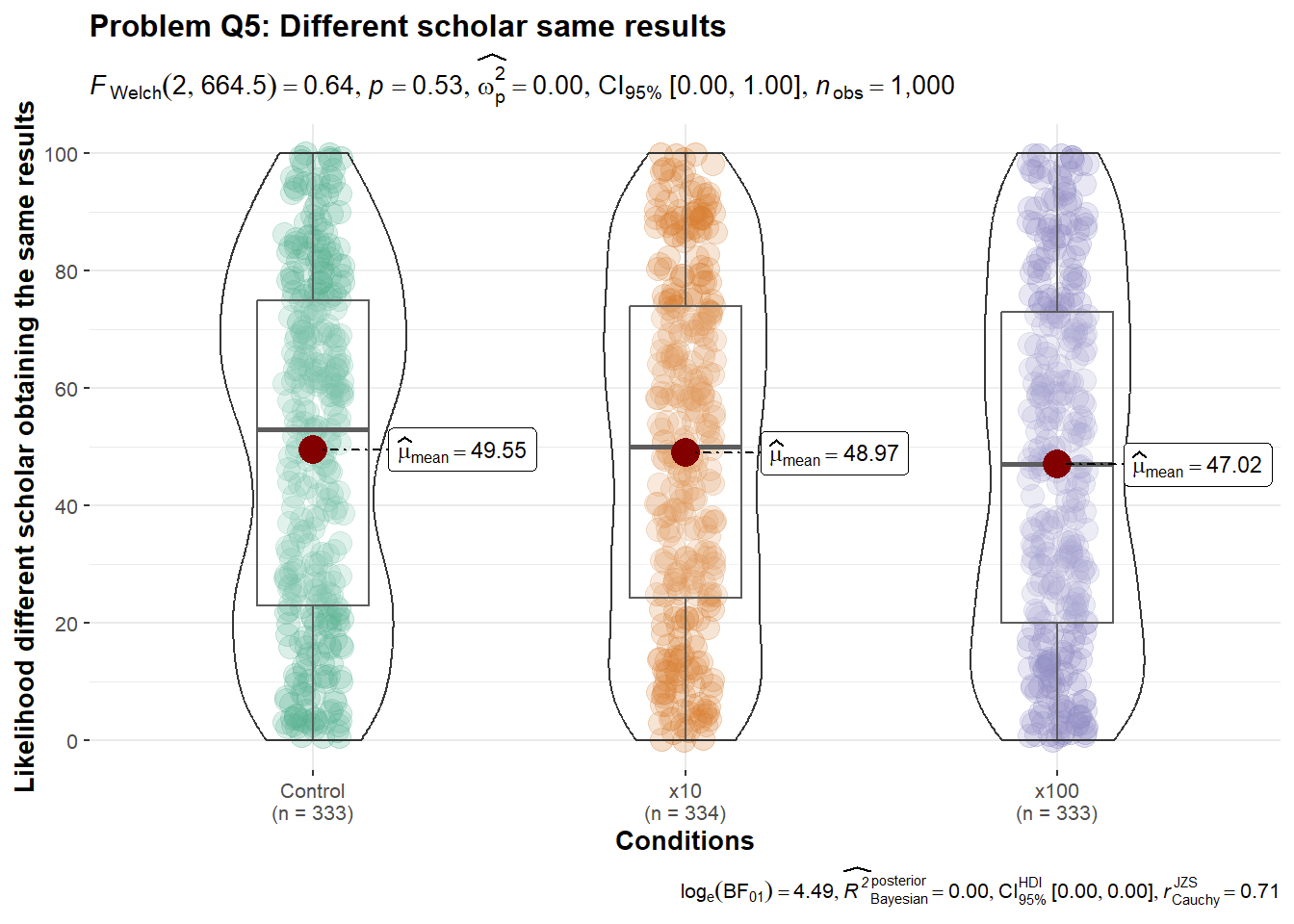
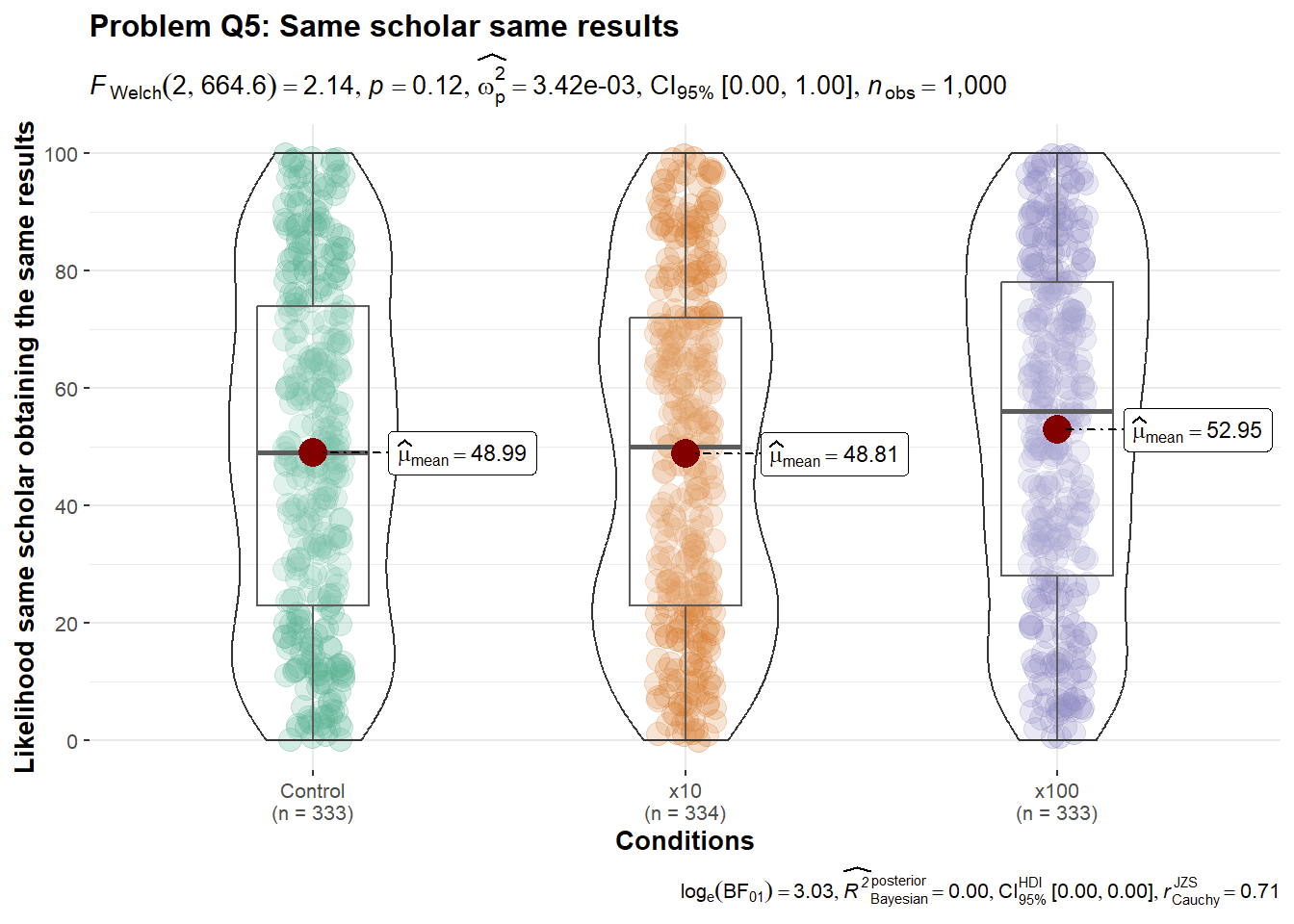


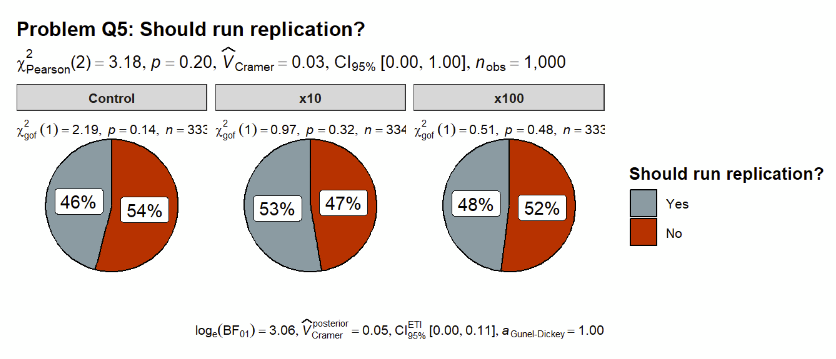


**Figure 1e**

*Problem Q5 “Exploratory analyses”*

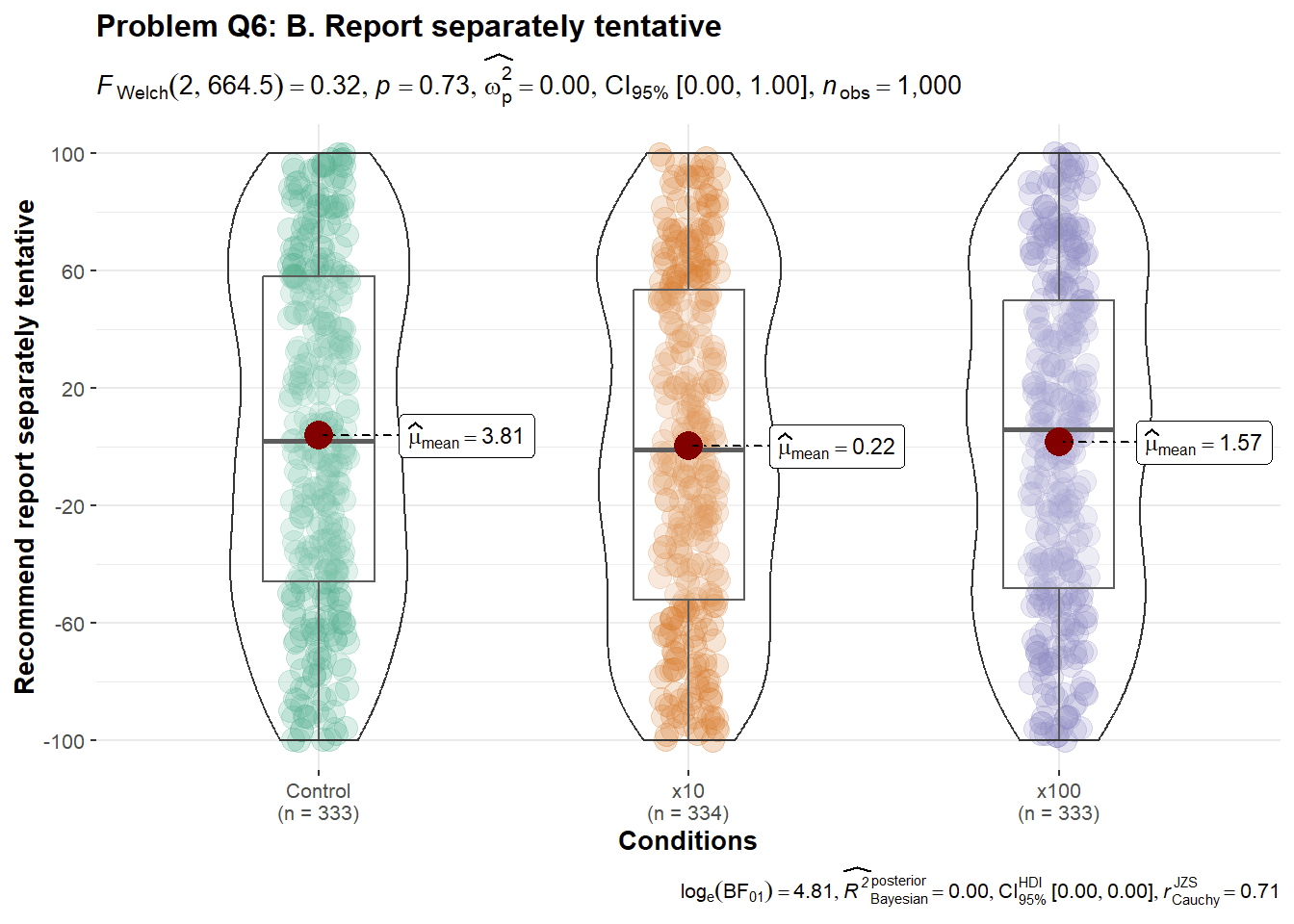
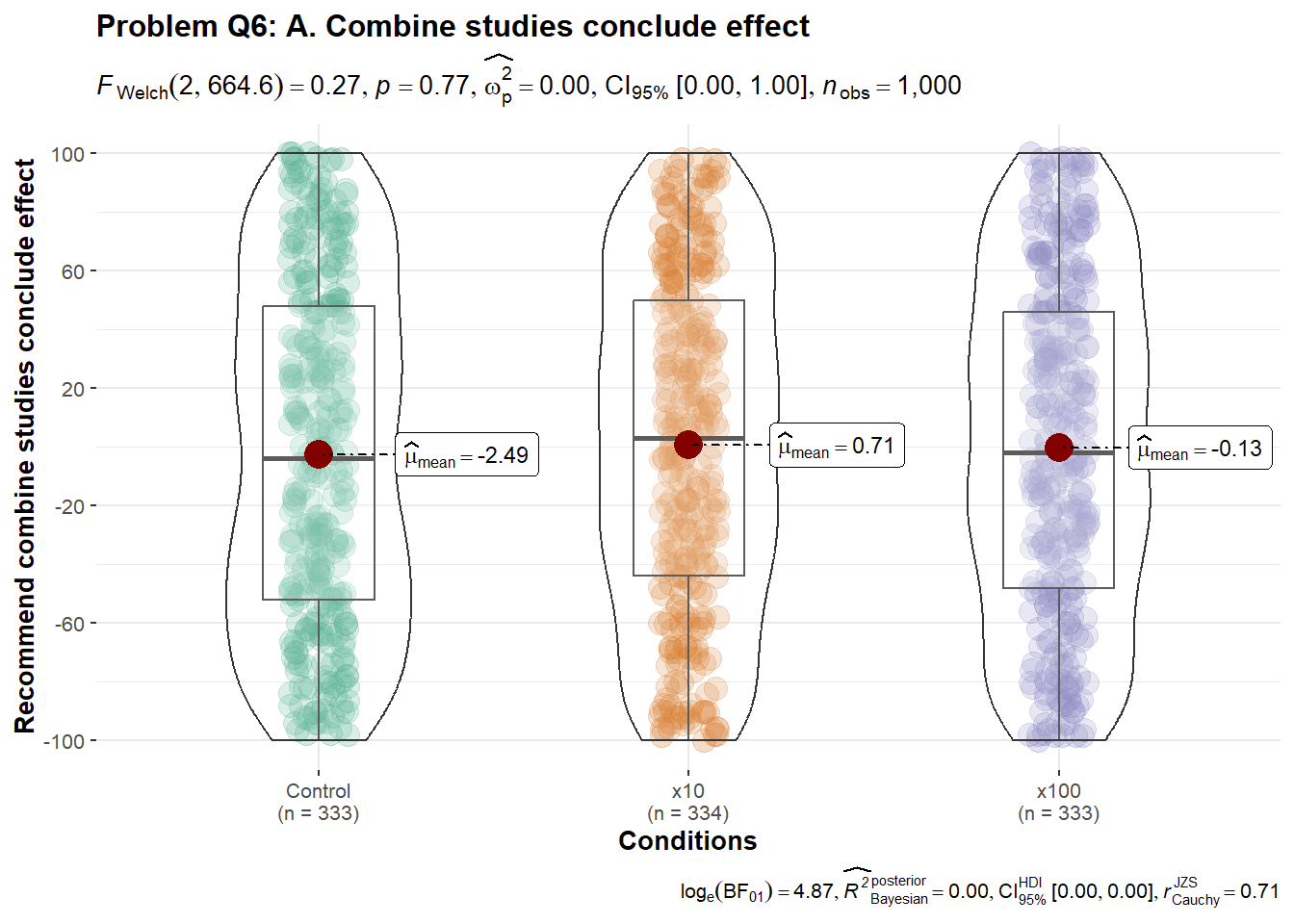


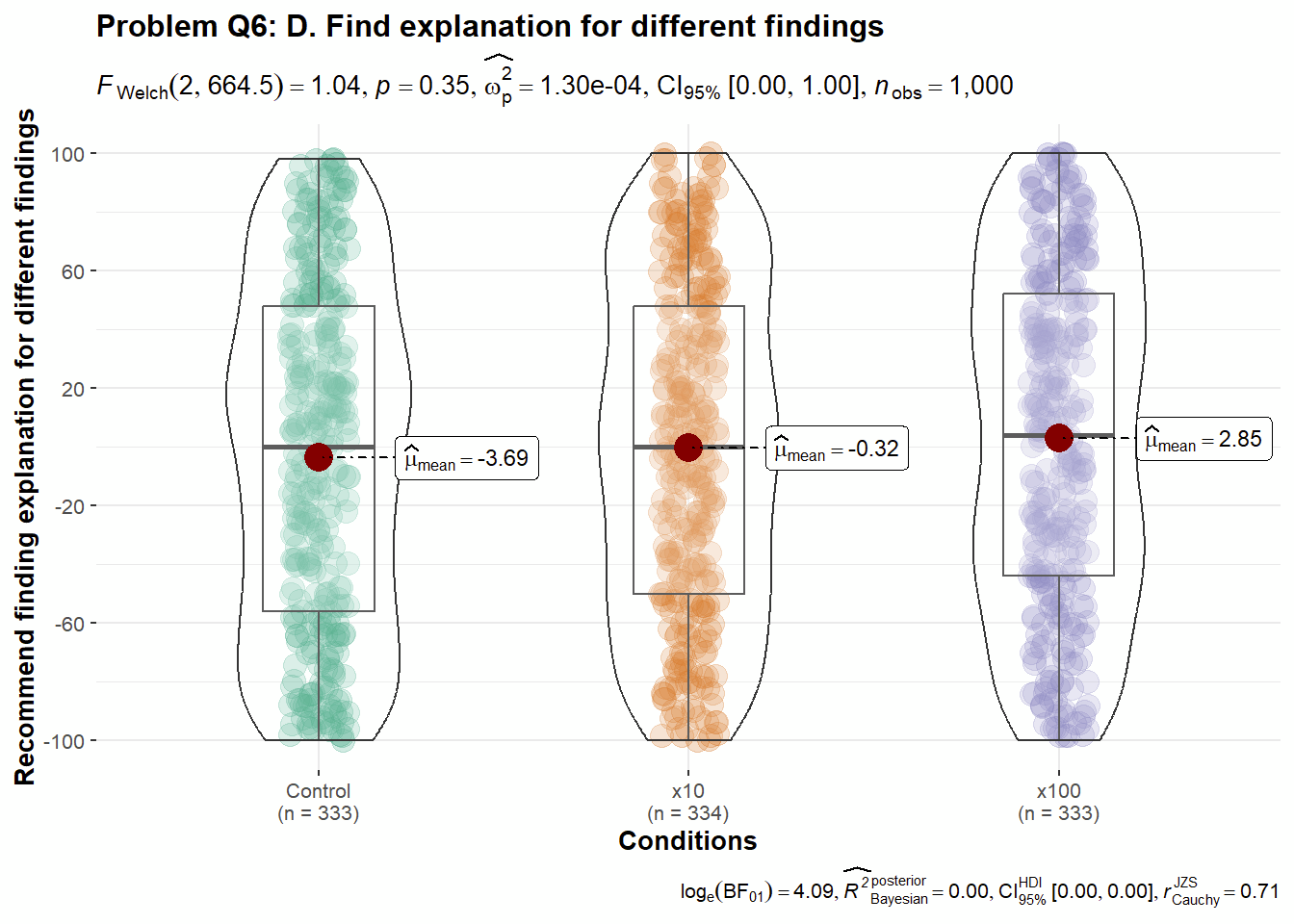
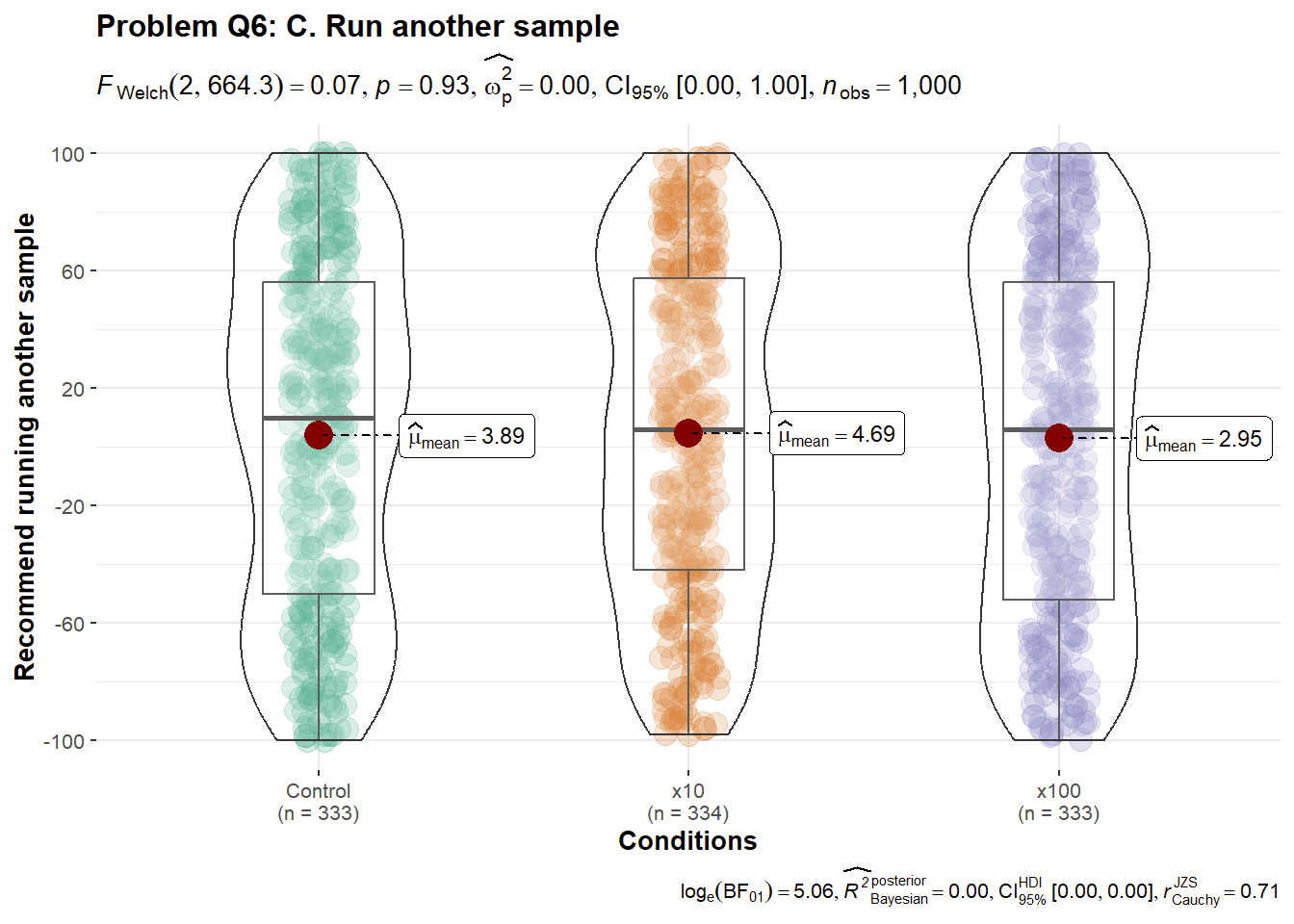




**Figure 1f**

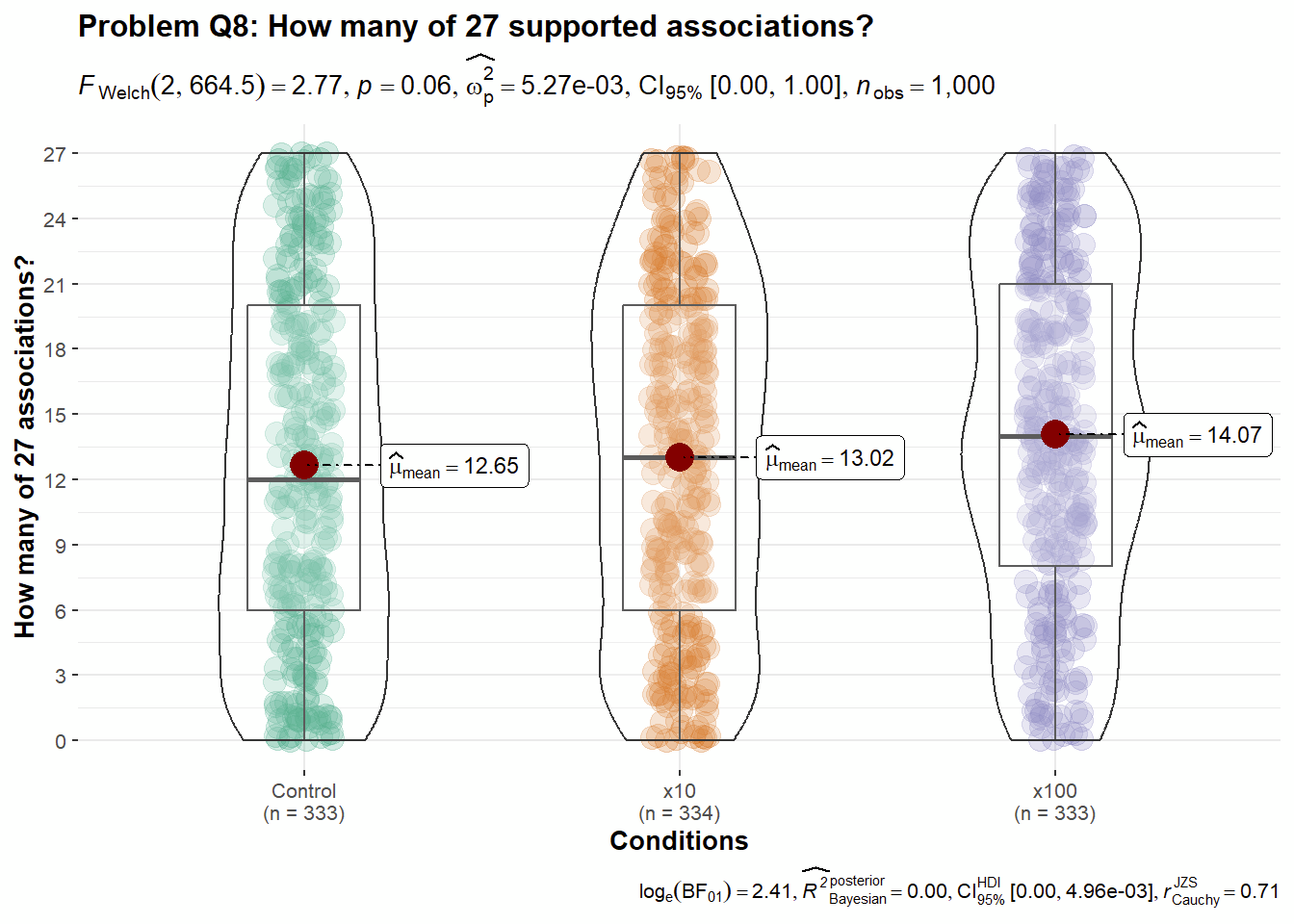
*Problem Q6 “Failed replication”*

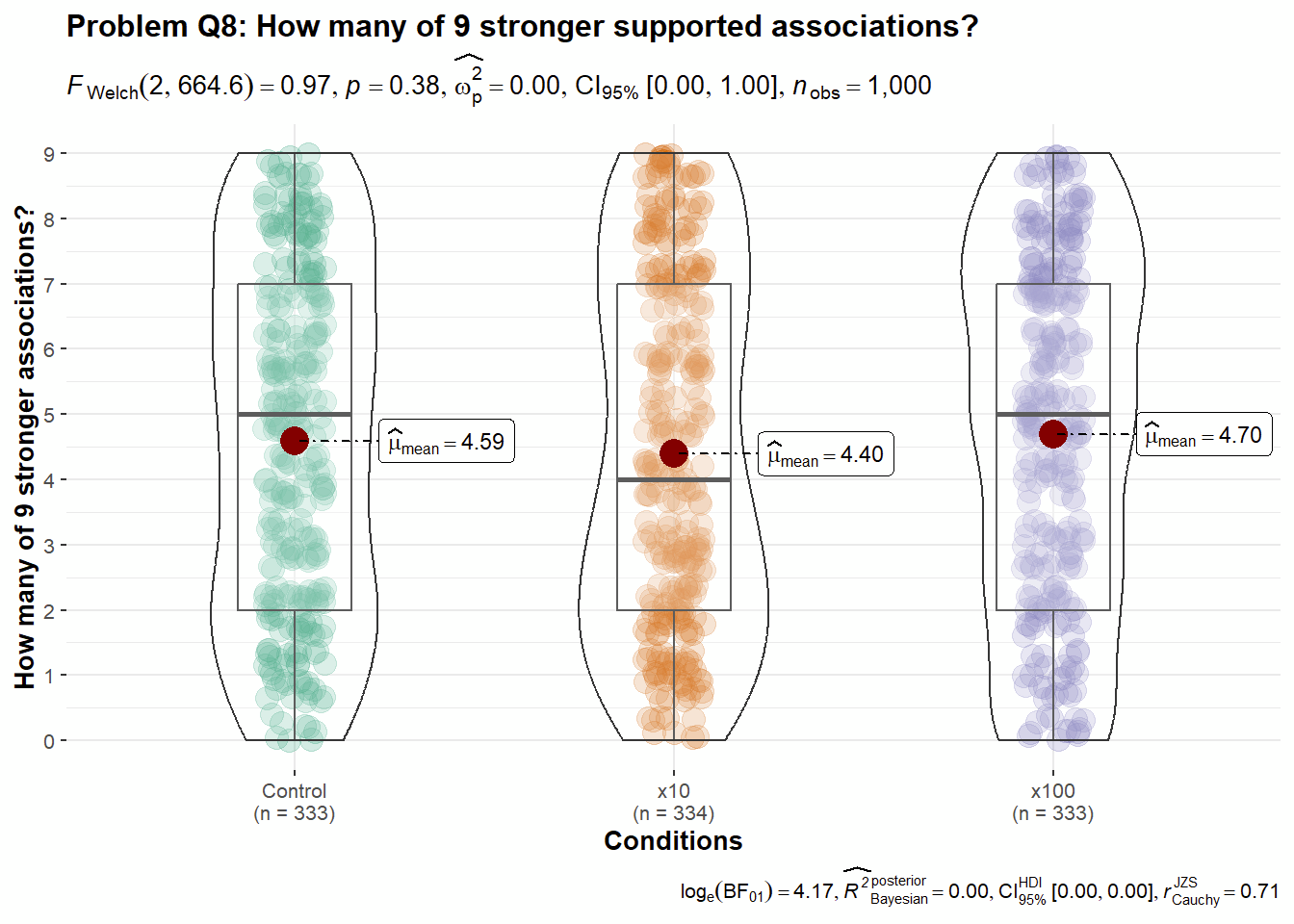




**Figure 1g**

*Problem Q8 “20 variables, 190 correlations”*





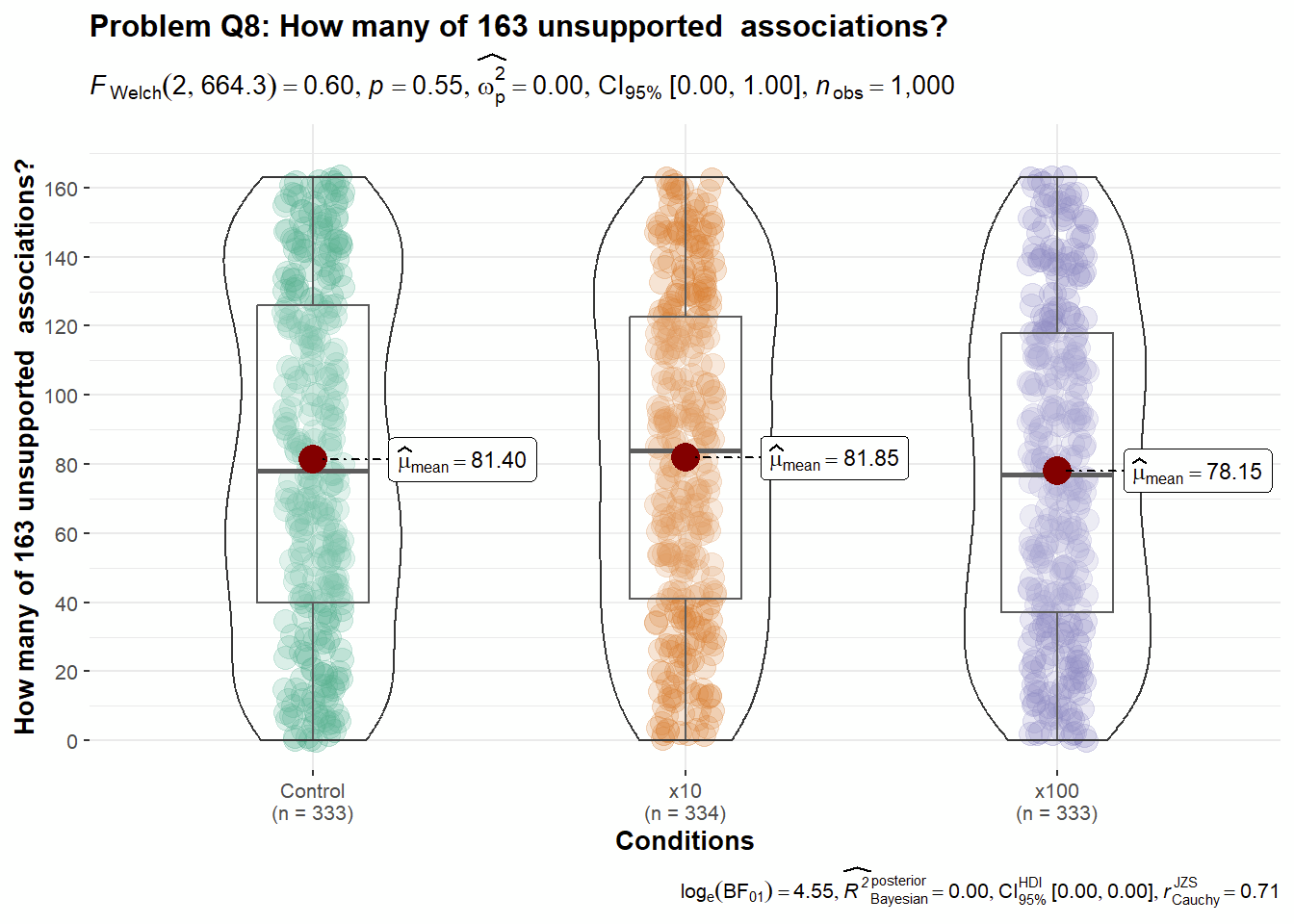


Table 7



*Summary of statistical tests for replication measures:*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Question |  | t-stat | df | *p* | Mean difference | Cohen's *d* and CI | Interpretation |
| 1 | DV 1 |  |  |  |  |  |  |
| 2 | DV 1 |  |  |  |  |  |  |
| 3 | DV 2 |  |  |  |  |  |  |
| 4 | DV 1 |  |  |  |  |  |  |
| 5 | DV 4 |  |  |  |  |  |  |
| 6 | DV 1A |  |  |  |  |  |  |
| DV 1B |  |  |  |  |  |  |
| DV 1C |  |  |  |  |  |  |
| DV 1D |  |  |  |  |  |  |
| DV 1E |  |  |  |  |  |  |
| 6 | DV 2 |  |  |  |  |  |  |
| 8 | DV 1 |  |  |  |  |  |  |

*Note*. One sample t-test, *N* = 1000. CI = 95% confidence intervals. The interpretation of outcome is based on LeBel et al. (2019).

Table 8

*Summary of ANOVA analysis for replication DVs*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Problem | DV | df | Mean Squares | *F* | *p* | *η2* | *Confidence intervals* |
| 1 | DV 1 |  |  |  |  |  |  |
| 2 | DV 1 |  |  |  |  |  |  |
| 3 | DV 2 |  |  |  |  |  |  |
| 3 | DV 3 |  |  |  |  |  |  |
| 4 | DV 1 |  |  |  |  |  |  |
| 4 | DV 3 |  |  |  |  |  |  |
| 4 | DV 4 |  |  |  |  |  |  |
| 4 | DV 5 |  |  |  |  |  |  |
| 4 | DV 6 |  |  |  |  |  |  |
| 4 | DV 7 |  |  |  |  |  |  |
| 5 | DV 1 |  |  |  |  |  |  |
| 5 | DV 2 |  |  |  |  |  |  |
| 5 | DV 4 |  |  |  |  |  |  |
| 5 | DV 5 |  |  |  |  |  |  |
| 6 | DV 1A |  |  |  |  |  |  |
| 6 | DV 1B |  |  |  |  |  |  |
| 6 | DV 1C |  |  |  |  |  |  |
| 6 | DV 1D |  |  |  |  |  |  |
| 6 | DV 1E |  |  |  |  |  |  |
| 6 | DV 2 |  |  |  |  |  |  |
| 6 | DV 3 |  |  |  |  |  |  |
| 8 | DV 1 |  |  |  |  |  |  |
| 8 | DV 2 |  |  |  |  |  |  |
| 8 | DV 3 |  |  |  |  |  |  |
| 8 | DV 4 |  |  |  |  |  |  |

*Note.*

Table 8

*Summary of post hoc analysis for replication DVs*

| Question | Variable | Comparison | | Mean difference | SE | df | *ptukey* | Uncorrected *p* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | DV1 Replication likelihood | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 2 | DV1 Estimated mean IQ | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 3 | DV2 Replication likelihood | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 3 | DV3 Required sample size | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 4 | DV1: Likelihood with estimated sample size | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 4 | DV3: Likelihood of correlation equal population | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 4 | DV4: Likelihood of correlation equal larger than population | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 4 | DV5: Likelihood of correlation larger than population | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 4 | DV6: Likelihood of correlation equal smaller than population | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 4 | DV7: Likelihood of correlation smaller than population | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 5 | DV1: Replication likelihood if same person reran | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 5 | DV2: Replication likelihood if someone else reran | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 5 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 5 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 | DV 1A | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 | DV 1B | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 | DV 1C | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 | DV 1D | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 | DV 1E | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 6 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 8 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 8 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |
| 8 |  | Control | X10 |  |  |  |  |  |
| Control | X100 |  |  |  |  |  |
| X10 | X100 |  |  |  |  |  |

Table 9

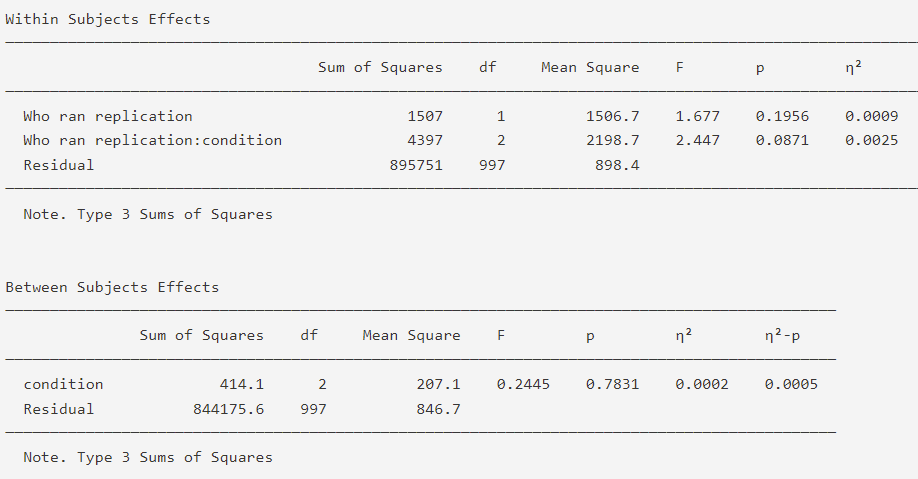
*Descriptive statistics of chi square analysis*

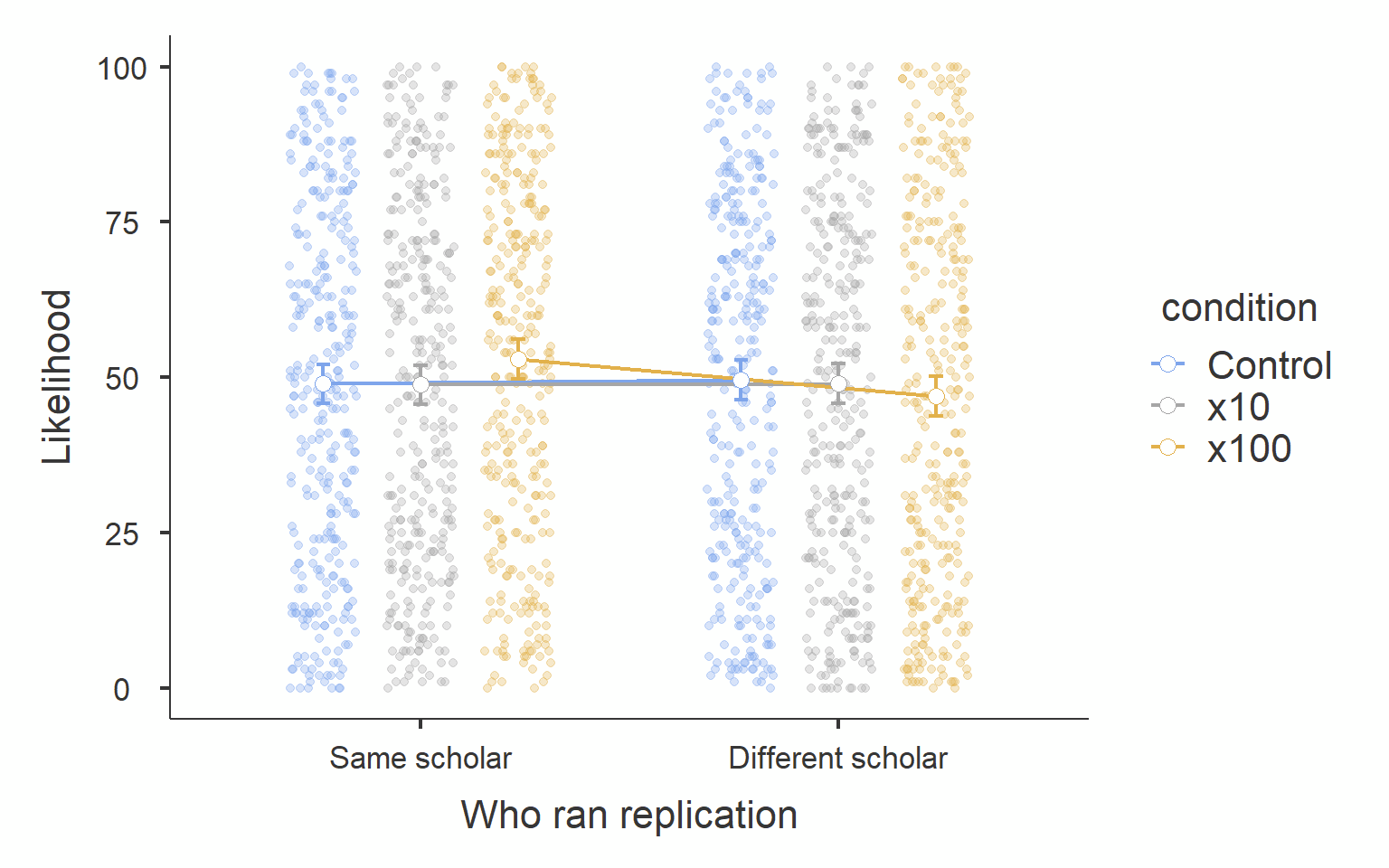
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Question | Response | Control | X10 | X100 | Chi square tests of independence |
| Q3: “Infants”  DV1 Rejection of null hypothesis (replication) | Yes | 91 | 90 | 110 | *X2* (2) = 5.68, *p* = .058  *N* = 514 |
| No | 80 | 81 | 62 |
| Q5:  “Exploratory analyses”  DV3: Should the study be rerun (replication) | Yes | 88 | 92 | 80 | *X2* (2) = 1.80, *p* = .407  *N* = 493 |
| No | 77 | 72 | 84 |
|  |  |  |  |  |  |

## Extensions

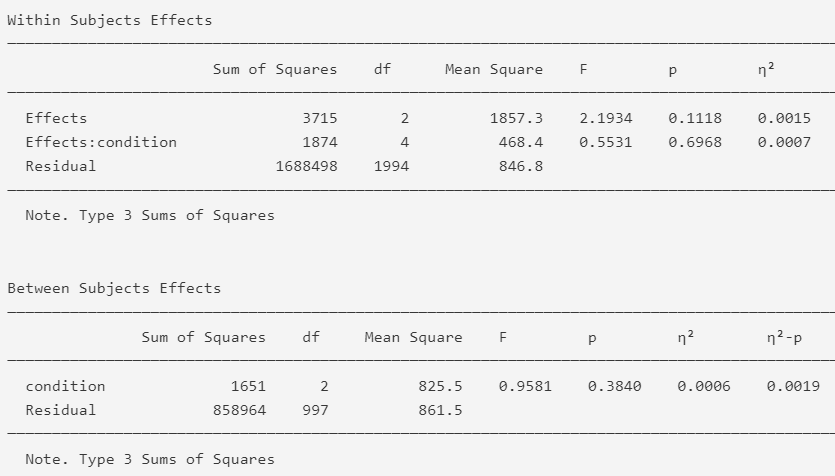
[See data analysis strategy “Extensions” and method’s “Exploratory extensions”, to be updated in Stage 2. ]

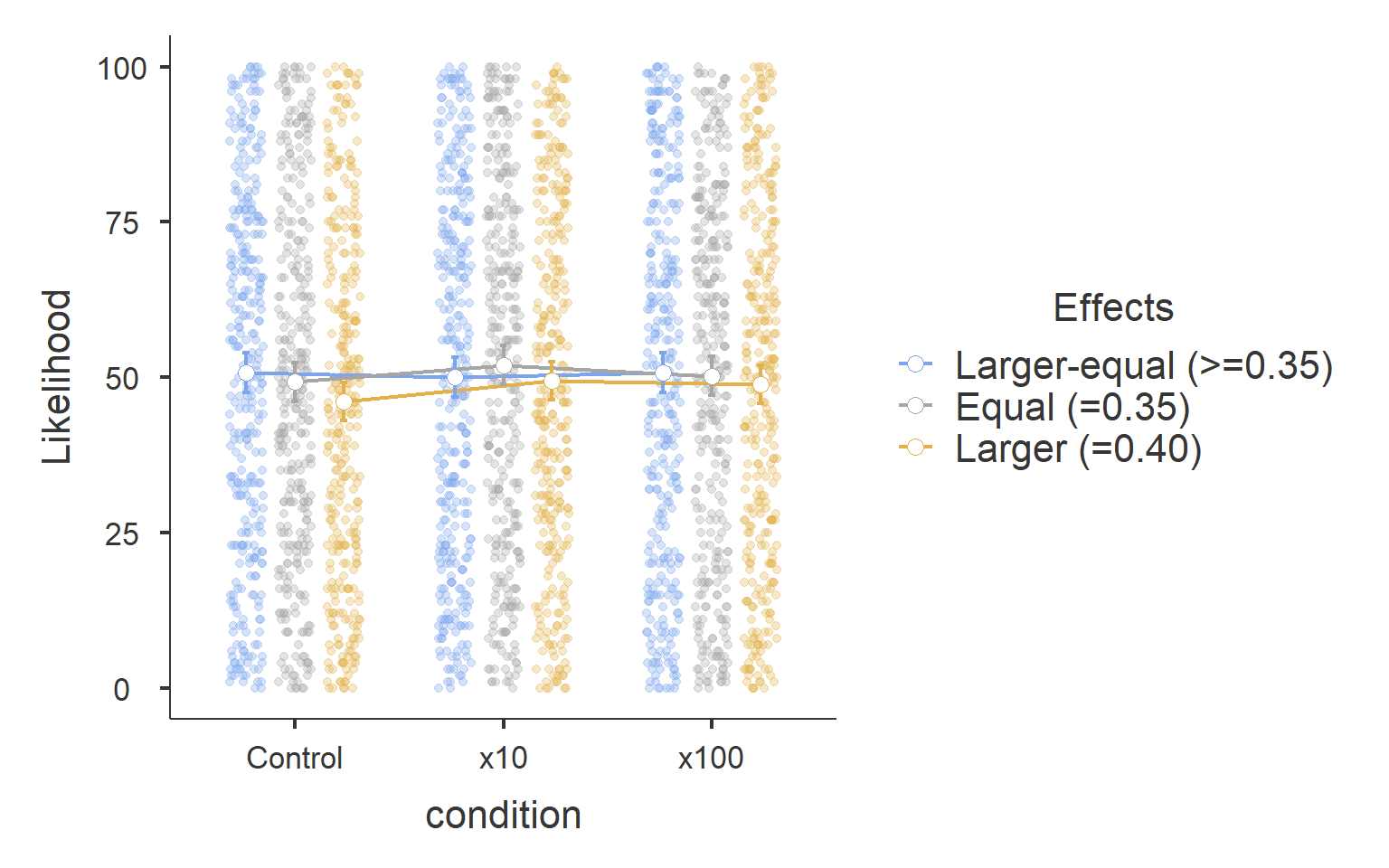
[Example for contrasts of same versus different scholar:]





[Example for likelihood of effects analyses:]





## Robustness checks

[See data analysis strategy “Robustness checks”, to be updated in Stage 2.]

## Exploratory analyses

[This section will be updated in Stage 2.]

## Winsorizing outliers

[In case of failed replication (less than 50% of replication findings are supported, we will conduct additional analyses with winsorizing outliers. See data analysis strategy section. ]

## Comparing replication to original findings

[We will compare answers from DVs which the original article had provided answers for: Problem Q1 DV1, Problem Q2 DV1, Problem Q3 DV1, Problem Q4 DV1, Problem Q5 DV3 and DV4, Problem Q6 DV1, and Problem Q8 DV1.]

# Discussion

## Exploratory analyses

[Please note that the exploratory analyses are only to be completed in Stage 2 following data collection]

## Limitations and future directions

[Please note that the discussion is only to be completed in Stage 2 following data collection]

[Planned discussion: Manipulating original and replication numbers separately. As we explained in the methods: “We note that in the scenarios we manipulated all the mentioned sample sizes in the question. For example, in Problem Q1, we varied both the sample size of the described original experiment of 20 (to 200 and 2000), and the sample size of the described replication of 10 (to 100, and 1000). An alternative approach could have been to only vary the described original experiment or to only vary the described replication. We decided to vary both because we wanted to keep the ratio between the original and the replication constant, to be able to examine the overall use of sample size information.” and a direction for future research could be to vary both parameters. We will discuss possible insights that can be gained from such an experiment.]

[Planned discussion of a direction for future research based on the suggestion by reviewer Dr./Prof. Romain Espinosa: To add conditions that not only focus on increasing sample size by 10 or 100, but also decrease sample size. For example, his suggestion was that in Problem Q2 one could add a condition of a sample smaller than 50.]

[Planned discussion raised by reviewer Dr./Prof. Kariyushi Rao: Limitation in the target study about the questions being about self versus others’, which may incorporate some degree of self-efficacy. In the laypersons version everything had to be translated to others’, and therefore we indirectly addressed this issue.]

[Planned discussion raised by reviewer Dr./Prof. Kariyushi Rao: Can run additional studies comparing the use of the original scenario, to our scholars adjusted scenario, to our laypersons scenario, for both scholars and laypersons, to see how the framing affects interpretation.]

[Planned discussion regarding robustness checks, implications, and limitations.]

[Planned discussion: Implications for the results for the debate regarding “belief in the law of small numbers” versus the “empirical law of large numbers” (Sedlmeier & Gigerenzer, 1997), especially in the context of the causal test in our extension]

# References

Bishop, D. V. M., Thompson, J., & Parker, A. J. (2022). Can we shift belief in the ‘Law of Small Numbers’?. *Royal Society Open Science*, *9*(3), 211028. <https://doi.org/10.1098/rsos.211028>

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333-342. <https://doi.org/10.1177/1745691614529796>

Chandrashekar, S., Cheng, Y. H., Fong, C. L., Leung, Y. C., Wong, Y. T., Cheng, B. L., & Feldman, G. (2021). Frequency estimation and semantic ambiguity do not eliminate‎ conjunction bias, when it‎ occurs: Replication and extension of‎ Mellers, Hertwig, and Kahneman (2001)‎. *Meta-Psychology*, *5*. <https://doi.org/10.15626/MP.2020.2474>

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal N-Pact Factors From 2011 to 2019: Evaluating the Quality of Social/Personality Journals With Respect to Sample Size and Statistical Power. Advances in Methods and Practices in Psychological Science, 5(4), 25152459221120217. doi: 10.1177/25152459221120217 <https://doi.org/10.1177/25152459221120217>

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*(10), e109019. <https://doi.org/10.1371/journal.pone.0109019>

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., ...Anvari, F. (2020). Use caution when applying behavioural science to policy. Nature Human Behaviour, 4(11), 1092–1094. doi: 10.1038/s41562-020-00990-w <https://doi.org/10.1038/s41562-020-00990-w>

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, *49*(2), 433-442. <https://doi.org/10.3758/s13428-016-0727-z>

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., (2017). Comment on “Reconstruction of a Train Wreck: How Priming Research Went off the Rails”. Retrieved from <https://replicationindex.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/comment-page-1/#comment-1454>

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269-275. <https://doi.org/10.1111/1467-9280.00350>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ...Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. Annual Review of Psychology, 73(1), 719–748. doi: 10.1146/annurev-psych-020821-114157 <https://doi.org/10.1146/annurev-psych-020821-114157>

Patil I., Makowski D., Ben-Shachar M., Wiernik B., Bacher E., & Lüdecke D. (2022). “datawizard: An R Package for Easy Data Preparation and Statistical Transformations.” *Journal of Open Source Software*, **7**(78), 4684. [doi:10.21105/joss.04684](https://doi.org/10.21105/joss.04684).

Piaget, J., & Inhelder, B. (1951/1975). The origin of the idea of chance in children.(Trans L. Leake et al).

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107-114. <https://doi.org/10.1177/2515245919838781>

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*(1), 33-51. <https://doi.org/10.1002/(SICI)1099-0771(199703)10:1%3C33::AID-BDM244%3E3.0.CO;2-6>

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, *26*(5), 559-569. <https://doi.org/10.1177/0956797614567341>

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105. [https://doi.org/10.1037/h0031322](https://psycnet.apa.org/doi/10.1037/h0031322)

Yeung, S. K., & Feldman, G. (2022). Revisiting the Temporal Pattern of Regret in Action Versus Inaction: Replication of Gilovich and Medvec (1994) With Extensions Examining Responsibility. *Collabra: Psychology*, 8(1). <https://doi.org/10.1525/collabra.37122>

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature*, *490*, 7418. <https://doi.org/10.1038/nature.2012.11535>

Xiao, Q., Yeung, S. K., Dunleavy, D. J., Röseler, L., Elsherif, M., & Feldman, G. (2023) Effect sizes and confidence intervals guide. DOI: 10.17605/OSF.IO/D8C4G . Retrieved from: <https://osf.io/d8c4g/>

Zhan, S., & Savani, K. (2023). Relative insensitivity to sample sizes in judgments of frequency distributions: People are similarly confident in the results from 30 versus 3,000 observations. *Decision*, *10*(1), 61. [https://doi.org/10.1037/dec0000182](https://psycnet.apa.org/doi/10.1037/dec0000182)

Zhu, M. & Feldman. G. (2023). Revisiting the links between numeracy and decision making: Replication Registered Report of Peters et al. (2006) with an extension examining confidence. *Collabra:Psychology*. Retrieved from <https://osf.io/62wqb>. DOI: 10.17605/OSF.IO/4HJCK