ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
Cognition and decision making, Psychology and cognitive neuroscience, Statistics

Keywords:
reasoning, statistics, statistical cognition, decision making, significance testing

Author for correspondence:
Richard D. Morey
e-mail:
moreyr@cardiff.ac.uk

Use of significance test logic by scientists in a novel reasoning task

Richard D. Morey[1], Rink Hoekstra[2]

[1]School of Psychology, Cardiff University
[2]Faculty of Behavioural and Social Sciences, University of Groningen

Although statistical significance testing is one of the most widely-used techniques across science, previous research has suggested that scientists have a poor understanding of how it works. If scientists misunderstand one of their primary inferential tools the implications are dramatic: potentially unchecked, unjustified conclusions and wasted resources. Scientists' apparent difficulties with significance testing have led to calls for its abandonment or increased reliance on alternative tools, which would represent a substantial, untested, shift in scientific practice. However, if scientists' understanding of significance testing is truly as poor as thought, one could argue such drastic action is required. We present evidence using a novel reasoning task that scientists may understand the logic of significance testing better than previously thought. Scientists may not be as statistically-challenged as often believed; reforms should take this into account.

THE ROYAL SOCIETY
PUBLISHING

For most of the past century, the dominant method of statistical inference has been statistical significance testing (SST). In a significance test, the statistical evidence in the form of a test statistic is compared to what would be expected under a particular hypothesis (often called the "null" hypothesis). If it would be surprising to observe evidence as strong as what was observed under this hypothesis, the evidence is deemed strong enough to call the assumed hypothesis into doubt, at least tentatively [see also 1,2]. The rarity of evidence as strong as what was observed under the assumed hypothesis—the so-called p value—is the typical way that results of significance tests are reported. The key feature of SST for our purposes is the assessment of evidence by means of comparing a result to a "null" distribution.

Despite the use of SST in a majority of research projects across fields, there is debate over whether scientists understand SST and can use it competently. Methodologists and statistical cognition researchers point to evidence from questionnaires and vignette studies to argue that researchers do not, in fact, grasp the core logic of SST. In one highly influential study of research psychologists, Oakes [3] presented six statements about a hypothetical significance test result to be categorized as true or false (e.g., "[The p value provides] the probability of the null hypothesis being true"). Despite all of these statements being false, 97% of the research psychologists categorized at least one as true. Oakes argues that this shows that the participants have an "[un]sound understanding of the logic of the significance test" (p. 82).

Oakes' basic method and results have been replicated and extended with various groups, showing that students [4], instructors [5], and statisticians [6] all misinterpret SST results. Moreover, these misinterpretations are difficult to eliminate even through targeted interventions [7]. As a result, many have argued that use of SST should be discontinued or dramatically reduced, and may even contribute to wide-spread replication problems in the sciences [3,8–11].

The interpretation of studies of researchers' understanding of SST is limited, however, by their methodology. A typical study presents a vignette describing research results. Statistical results are offered to the participants (e.g., a t statistic and p value), who are then asked to explicitly give or endorse various interpretations. These responses are taken to represent participants' understanding, or misunderstanding, of SST. However, there are reasons to be cautious of drawing strong conclusions from these studies, including the abstract nature of such vignettes, the lack of investment researchers have in the fictional research, and their disconnection from research activity (e.g., experimentation and replication). It is unclear how well vignette studies (including ones by the present authors: [12,13]) tap understanding of the core logic of SST rather than, say, familiarity with the technical terminology used to present statistical results. Conceptual understanding and fluency with common representations are both important, but are distinct.

A second major piece of evidence for misunderstandings of SST logic is reasoning errors in published papers [14–17]. Like evidence from vignette studies, however, these errors are difficult to interpret as misunderstandings of SST logic *per se*. These examples show that whatever process lead to the statistical conclusion was flawed in some way, but many processes contribute to such conclusions. Cognitive [e.g., 18], technological [e.g., 19], and social processes [e.g., 20] have all been assigned some blame for statistical reasoning failures.

In deciding how to improve statistical reasoning, it is crucial to know where the problems lie. The *fact* of reasoning problems tells us little about their *source*. In assessing potential interventions, however, the source is crucial. Some interventions might focus on the social aspects (e.g., decreasing the need for "significant" results for prestige), some on technological aspects (e.g., presenting statistical results in ways that were previously impossible), and some on cognitive aspects (e.g., adopting Bayesian procedures because these are claimed to be better understood).

To avoid conflating basic reasoning failures and lack of fluency with common statistical terminology, we avoid using common statistics—or, indeed, any numbers—at all. Instead of focusing on familiar statistical language and tests participants' fluency with existing procedures, we adopt a different approach: we test working scientists' understanding of the basic conceptual framework underlying SST using a simulated experimental task.

The key innovation allowing us to focus on SST reasoning was to design an experiment that prevents the use of alternative strategies. A critical feature of SST is that the use of a null distribution destroys information about effect sizes and sample size. In fact, this aspect of SST reasoning is often criticized, while alternative methodologies focus on effect sizes (point estimates, confidence intervals, equivalence, likelihood, Bayesian priors/posteriors). We offered our participants only the information in a p value, and participants had to understand or discover how to obtain that information. Their task was to use this information to come to a decision about the true sign of an effect through repeated experimentation.

If participants have poor understanding of SST, they would 1) often come to the wrong conclusion, in spite of ample information; 2) show error rates that are only weakly associated with true effect size; 3) be unable to articulate strategies for performing our task; 4) be sensitive to misleading, task-irrelevant information; 5) be insensitive to SST-relevant information. The scientists in our sample often came to the right conclusion, and their performance showing sensitivity to the SST-relevant information they were given. Moreover, they explicitly reported using SST strategies. Our results suggest that common methods for assessing scientists' competence may miss important aspects of their statistical knowledge, and hence that the case for abandoning significance testing may be overstated.

1. Testing reasoning by withholding information

In tests of perception, it is common to eliminate one cue in order assess the ability to use another: e.g., eliminating brightness cues to test colorblindness [21]. If color is the only useful cue for reading a number on a card, deficits in color vision make the number difficult to read. We adopt a similar strategy to test statistical reasoning: we eliminate numerical information from statistical results to test scientists' ability to interpret results with reference to a null sampling distribution, a central element of SST logic. Without numerical information, many other strategies and heuristics, such as confidence intervals, or Bayesian inference, are difficult or impossible to apply.[1]

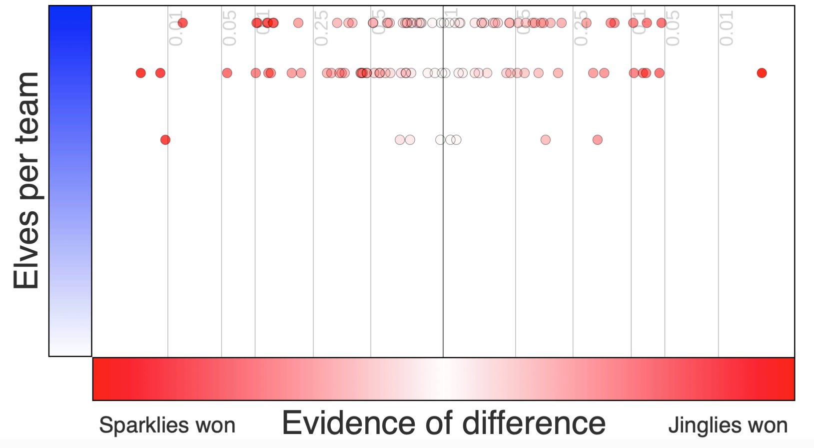
Participants were scientists or trainees recruited via social media. Our statistical reasoning task required them to perform a series of experiments to judge which of two groups of "Christmas elves" — "Jinglies" or "Sparklies" — could make more of a particular toy. A demonstration version of the task can be found at `https://richarddmorey.github.io/Morey_Hoekstra_StatCognition/articles/task_demo.html`. Because the study was run around the Christmas holiday season, we hoped the theme would make the task more engaging. The numerical information for an experiment, including sample size and the test statistic, was translated into color and location and displayed as a point on a two-dimensional visual interface (Figure 1). Participants could change the sample size per group for each experiment (increasing the time required to return a result), but did not know its numerical value. Importantly, the meaning of the colors and locations was unknown to the participants, aside from the monotone relationship with the sample size and statistical evidence.

Participants were randomly assigned to one of 15 effect size conditions: either no difference ($\delta = 0$), or $\delta = \pm 0.1, \pm 0.185, \pm 0.296, \pm 0.433, \pm 0.596, \pm 0.785$, or ± 1 standard deviation units. Each participant had a 25% probability of being assigned $\delta = 0$, with the other 75% being randomly and uniformly distributed across the remaining 14 effect size conditions. These true effect sizes were not revealed to the participants. Their goal was to determine the sign of the effect (i.e., which of the two teams is truly faster).

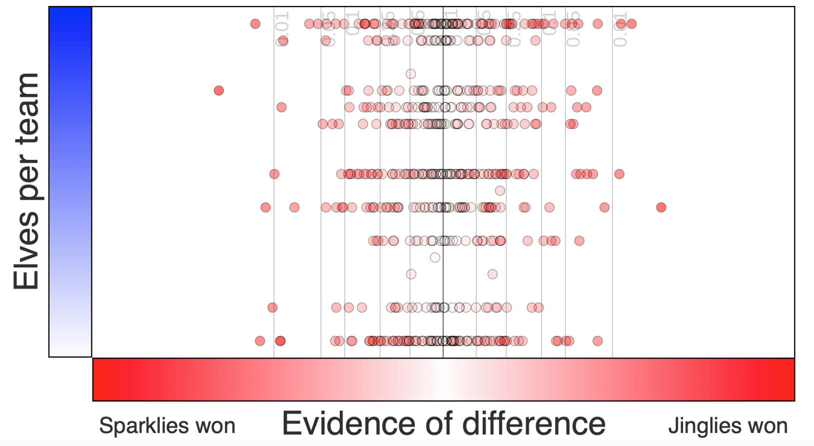
Consistent with the fictional two-sample design, statistical evidence for each "experiment" was sampled from a normal distribution with mean that depended on the (chosen, but unknown) sample size and their randomly assigned effect size:

$$Z \sim \text{Normal}(\delta\sqrt{n/2}, 1)$$

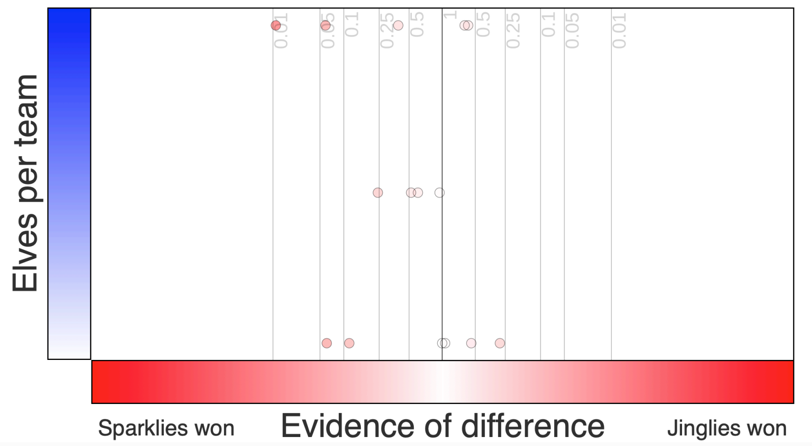
[1] A formal statistical explanation showing that the task is difficult or impossible to perform using non-SST logic is given in Section 3 of Supplement A.



(A) Random shuffle reports by one participant in the "wide" condition.



(B) Random shuffle reports by one participant in the "narrow" condition.



(C) Experimental samples by the same participant as shown in (B). This participant responded that the two groups were the same; the true effect size was -0.1, so their response was a false negative.

Figure 1: Examples of the experimental interface with several participants' samples. The x-axis monotonically (but nonlinearly) related to the strength of the statistical evidence (z statistic) favoring one group; the y-axis is monotonically (but nonlinearly) related to the sample size. Underlying numerical values of the statistical evidence and sample sizes were unknown to the participant. Corresponding p values and vertical lines are given for reference; they were not shown to the participants.

The Z test statistic was mapped into a horizontal location on the interface through an arbitrary function unknown to the participant. Participants were randomly assigned to one of two mapping functions: a "wide" function, and a narrow function (see Supplement A, section 1.2 for full mathematical details). Figures 1A and B show the visual effect of this manipulation. Statistically, these two conditions were identical; visually, they were not.

This visual manipulation was crucial to study, because it allows assessment of participants' use of the null sampling distributions. In addition to being able to sample fictional "experiments", participants could sample "random shuffle reports" that were described as the results of experiments with random assignment of elves to groups: that is, the result of experiments in which the null hypothesis was true. These results took no time to return. Participants were not told how to use these samples, only that they might use them.

Our experiment was constructed such that the only way to assess the evidence in the data was by comparison of the fictional experimental results to a null sampling distribution: either the one provided by the random shuffle reports, or a simpler null that assumes that the evidence will favor one team or the other with 50% probability. Thus, the information afforded only the information in a p value, but it was not described as such; participants had to discover for themselves how to use the information.

After sampling as many "experiments" and "random shuffle reports" as they liked, participants could report whether they believed Jinglies or Sparklies were the better team, that they could not detect a difference, that there was no difference, or that they were bored and wanted to stop. Following their decision they were asked several open-ended questions about their strategy, along with some opinion and demographic questions. Our central questions are whether participants can effectively find the "truth", whether they report strategies consistent with SST, and whether their behaviour shows evidence of strategic SST use.

Here, we report the results of 506 scientists or trainees who completed the statistical reasoning task.

2. Participant sampling behavior

Participants sought out information that would be necessary for significance tests. They made heavy use of shuffle reports (Figure A2). Across all true effect sizes, participants sampled a median of 152 shuffle reports (range: 1-2034; in both panels A and B, lines show robust regression fits [22]).

Participants also made use of "replications" of the fictional experiments. Figure 2B shows the distribution of the number of experiments sampled as a function of the true effect size. Median numbers of experiments range from 20 when Jinglies and Sparklies were equally fast, down to 9 when the true effect size was $\delta = 1$ and thus the effect was relatively easy to detect ($Kruskal - Wallis\chi^2(7) = 45.70$, $p < .001$). When the effect size is small and difficult to detect, participants experimented more before deciding.

3. Success rates identifying effect sign

Decision rates as a function of true effect size are shown in Figure 3.

Of the 136 participants for whom the null hypothesis was true (i.e. $\delta = 0$), 20 participants (14.7%) incorrectly indicated an effect. This is larger than the typically-accepted 5% false positive rate in many sciences; however, participants were performing a novel task with no recourse to numbers or statistical software. Those who did not indicate an effect when $\delta = 0$ tended to indicate that they *did not detect* an effect (103; 75.7%), which is the correct conclusion from the SST perspective. The other 13 (9.6%) indicated that the groups were the same, which under SST is typically considered a fallacy.

When there was a true effect ($\delta \neq 0$), correct decisions increased as a function of effect size, plateauing at about 95%. Of the 370 participants for whom $\delta \neq 0$, only 2 (0.5%) indicated the incorrect team [a sign, or Type S, error; 23]. For larger effect sizes, participants never incorrectly indicated that the two groups were the same.

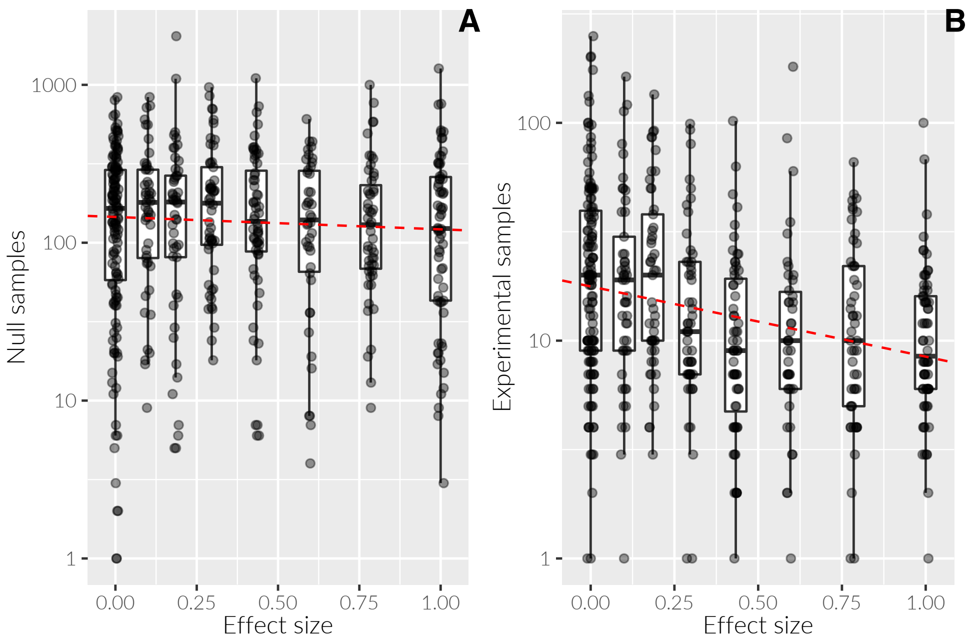


Figure 2: Sampling behavior by effect size. Each point represents a participant. A: Number of samples from the null distribution as a function of true effect size. B: Number of samples of fictional 'experiments' as a function of true effect size. Note that the y axis is logarithmically scaled. Lines are robust regression fits. Positive and negative effect sizes hav been collapsed.

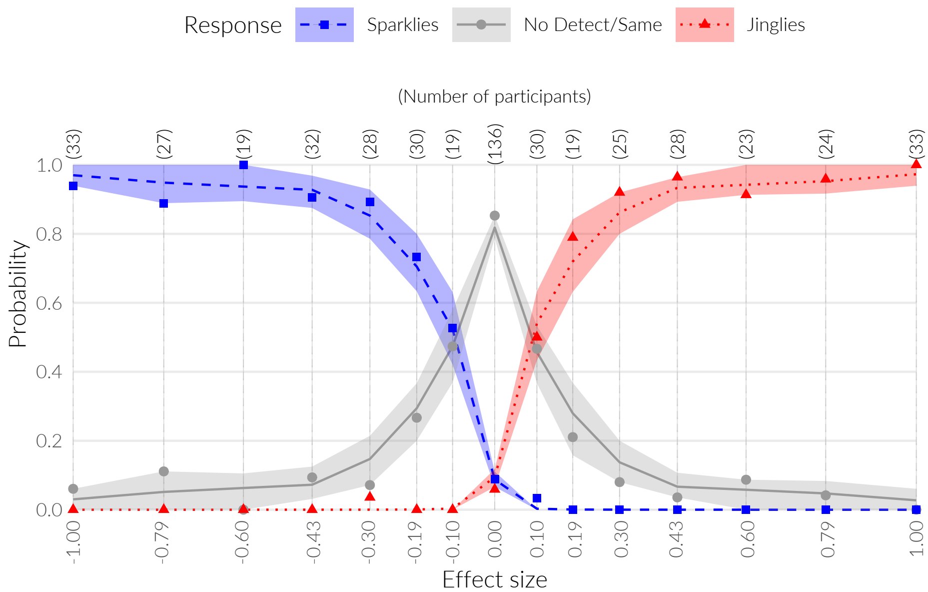


Figure 3: Observed and fitted probabilities for each effect size and response. For negative and positive effect sizes, the correct response "Sparklies" or "Jinglies" respectively. Fitted probabilities are from the signal detection model outlined in Supplement A. Lines show predicted probabilities; ribbons show where 68% of the observed probabilities should fall given the predicted probabilities. These limits are approximate due to the discreteness of the response.

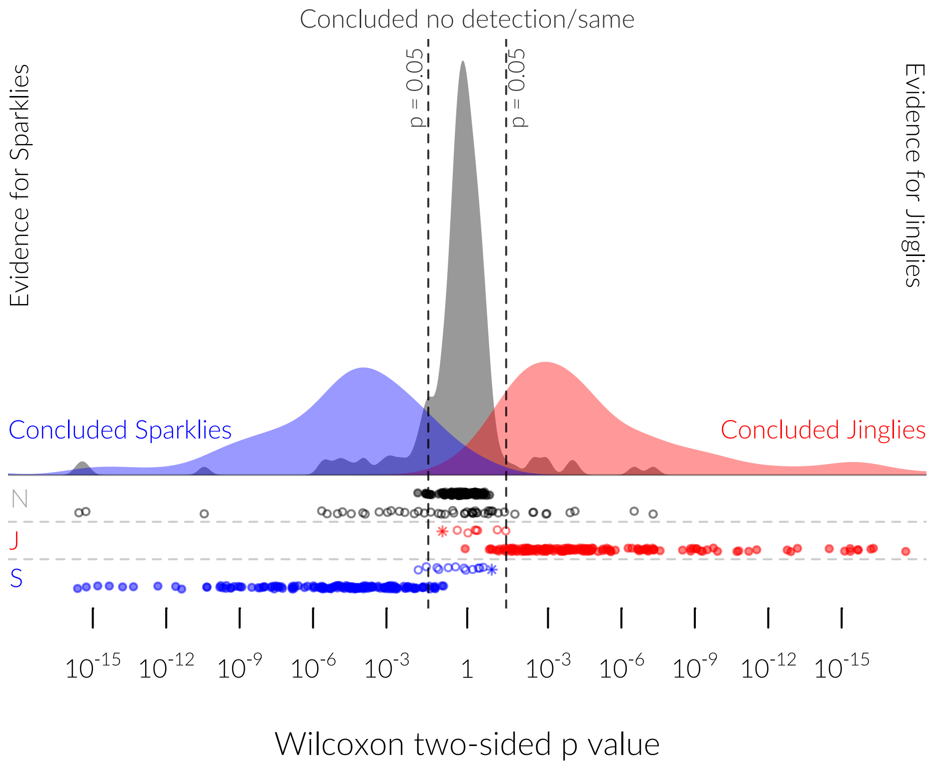


Figure 4: Statistical evidence underlying participants' decisions. The Wilcoxon p value (x axis) used as a rough index of evidential strength in the display. Kernel density estimates for the evidence are shown for three relevant conclusions. Each point at the bottom represents a single participant. Filled circles show correct decisions; hollow circles, incorrect decisions. The two asterisks show sign errors.

Signal detection theory gives us another perspective on the decision rate shown in Figure 3, allowing us to correct for the baseline of errors that occur in the null condition [24]. We combine the "false alarm" rate when $\delta = 0$ (14.7%) with the "hit rates" for all other conditions using a simple signal detection model; see Supplement A, section 8 for model details. The fitted model yields d' parameters that range from 1.41 when $\delta = 0.1$ to 3.24 when $\delta = 1$.

4. Use of information in the display

Another way of evaluating participants' responses is whether they reflect the information in the display at the time the decision is made, taking into account all points. To roughly quantify the evidence for a difference for each participant, we computed two p values from Wilcoxon tests using the fictitious experimental results as they stood when the participant made their decision: a signed-rank test on the experimental samples alone, and a rank-sum test between the shuffle reports and the experimental samples. These two p values indicate the information available to participants using sign-like significance tests and those using the null samples, respectively. The rank-sum p value is based on more information and so was typically lower. It makes little difference to the qualitative results, but to fairly account for the information available to the participant, we used the smaller of the two p values. In general, smaller p values suggest a larger observed between the shuffle reports and the experiments, allowing us to compare the stimulus the participants were given to their decisions.

Figure 4 shows the distribution of Wilcoxon p values (arranged by the direction of the decision). Kernel density estimates show the distributions of p values when participants indicated that Sparklies were faster, no detection/same, or that Jinglies were faster. With a few notable exceptions, participants' conclusions appear reasonable given the information in the display, though a few participants appear to ignore clear evidence of an effect. We provide an interactive app for exploring participants' individual responses at https://richarddmorey.shinyapps.io/explore/.

5. Sensitivity to SST-Relevant Information

In addition to a random effect size, participants were also randomly assigned to one of two transformations of the location/color test statistic from an underlying z statistic. Of particular interest was how the transformation affected responding for the same visual deviation from the center.

The visual effects of the manipulation are shown in Figure 1, panels A and B. The two experimental conditions used different arbitrary monotone mappings from the underlying Z-statistic to the visual space. Intuitively, this would be like deciding to use Z^3 instead of Z in all Z tests; one would need to adjust the significance criteria to account for the cubing (e.g., use $|1.96^3| = 7.53$ instead of $|1.96|$ for a $\alpha = 0.05$ level test), but the underlying test remains the same. The manipulation changes only the visual impression of the sampling distributions, allowing us to see how sensitive their responses are to the null sampling distribution as represented by the random shuffle reports.

If participants were using the shuffle reports to interpret the data, as would be predicted if they were using SST logic, the transformation should affect their interpretation of the visual evidence: a visually-extreme point should be more discounted against the sampling distribution that is wider. When we break down responses by the *visual* extremeness of the evidence, responses in two conditions should appear different; when we break down responses by *statistical* extremeness (i.e., the p value) responses in the two conditions should appear very similar, because the visual manipulation is irrelevant given the p value.

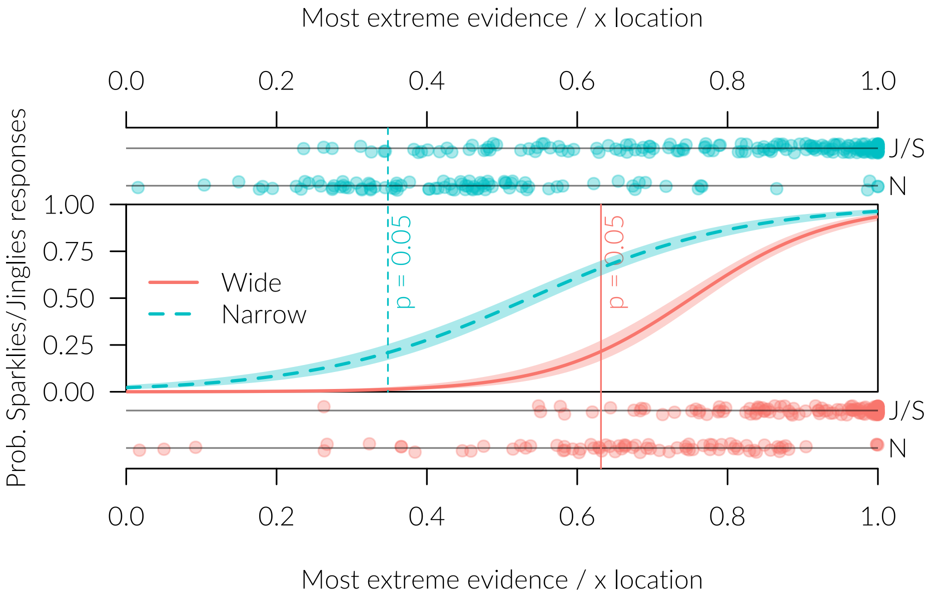
Figure 5 (top) shows responses (no detect/same or Jinglies/Sparklies) as a function of the most extreme experiment sampled (x axis) and the transformation. There was a strong effect of the transformation consistent with use of the null sampling distribution; participants randomly assigned to the "narrow" evidence transformation responded "Jinglies/Sparklies" for much less visually extreme evidence (sequential LRT: $\chi^2_2 = 35.492, p < .001$).

A logistic regression relating responses to the visual extremeness of the evidence and the transformation provides predicted probabilities of responding "Jinglies/Sparklies" when the visual evidence corresponded to $p = 0.05$ for the null sampling distribution. In both the wide and the narrow conditions, the predicted probability of a "Jinglies/Sparklies" response at the critical value was about 22%, despite that in the wide transformation condition this point was about twice as visually extreme.

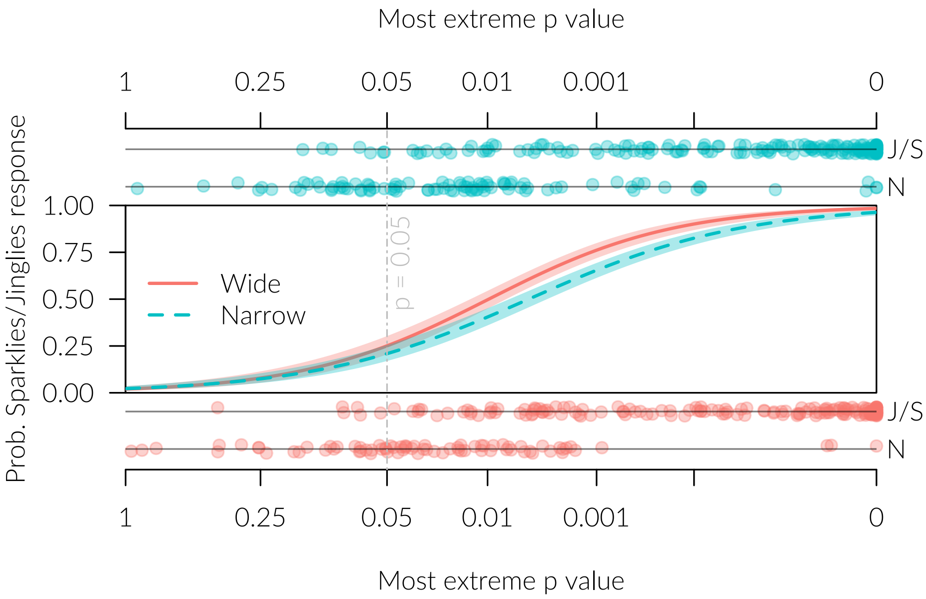
Applying the same analysis to the responses corrected for their respective sampling distributions (Figure 5, bottom) almost completely eliminates the effect of experimental condition, as would be expected if most participants were using the sampling distributions to calibrate (sequential LRT: $\chi^2_2 = 3.505, p = 0.173$). It is noteworthy that when the responses are aligned by sampling distribution, the wide condition appears to slightly dominate; this is consistent with some participants incorrectly using the non-diagnostic visual extremeness to perform the task. If more people had been fooled by the irrelevant width of the null sampling distribution, we would expect this effect to be substantially larger.

6. Self-Reported SST Strategies

After they reported their decision regarding which team they believed was faster, we asked participants three questions about how they performed the task: what was the most salient



(A) Predicted response probabilities relative to visual extremity. Vertical lines show the critical 0.05 for the corresponding null sampling distribution.



(B) Predicted response probabilities relative to the null sampling distributions (implicit p values).

Figure 5: The effect of the evidence transformation manipulation on responding. Points on top (narrow scale; $q = 7$) and bottom (wide scale; $q = 3$) represent participants' decisions as a function of the most extreme experiment sampled. See the methods details for the interpretation of q. "N" indicates a "no detect" or "same" response; "J/S" indicates a response in favor of a difference between the groups. Curves show predicted probability by a logistic regression fit with standard errors.

Table 1: Frequencies of self-reported strategies.

	Strong	Only weak	Neither	Total	No shuffles	Missing
Count	362	69	75	506	28	29
%	71.54%	13.64%	14.82%	100%	5.53%	5.73%

information for their decision, what was their general strategy, and whether/how they used the shuffle reports.

We coded their responses according to whether they indicated comparing to the shuffle reports or using them to assess sampling variability (which we term "strong" significance testing strategies), assessing asymmetry in the display (a "weak" significance testing strategy, because it ignores information), and whether they explicitly deny using the shuffle reports (see Supplement A, section 7 for coding details).

As Table 1 shows, a large majority of participants (362, 71.54%) indicated using strong significance testing strategies. We should be cautious in directly interpreting this high number alone, however, because participants were told in the instructions that the shuffle reports could be used for assessing sampling variability. We did this to make clear what the shuffle reports were, but without explaining *how* to use them. To some extent, then, the text responses may reflect the instructions. However, the data strongly suggest a deeper understanding; first, among the responses were richer, lucid descriptions of SST logic, such as:

> "[t]he random [shuffles] showed quite often such 'strong evidence', even at high sample sizes. That should not happen when the evidence is really strong, so probably the end of the scale was not [so] strong evidence... The random [shuffles] helped me to judge how common misleading evidence in that order of magnitude is, and after 5 samples from the real experiment I concluded that this result is probably not misleading evidence."

Secondly—and most importantly—the instructions did not tell the participants how they should use the shuffles reports, yet many participants gave detailed accounts. Combined with the other reported results, this strongly suggests that our participants—with some exceptions—do understand the basic SST logic and can deploy it to correctly solve novel problems.

7. Discussion

Although it has previously been suggested that scientists have dramatic misunderstandings of SST logic, scientists and trainees in our experiment demonstrate both understanding and the ability to use the logic to come to the correct conclusion in a simulated statistical task. Moreover, they report strategies consistent with SST, and signatures of SST reasoning can be seen in their responses. Because we removed numerical effect size and sample size information — making strategies other than pure significance testing difficult or impossible to apply — our results are evidence that scientists *can* successfully deploy SST logic. It is still an open question what causes typical SST reports to be misunderstood so often, but we have not found evidence that the problem is misapprehension of its underlying logic.

Our findings echo other demonstrations that human reasoning can, under some conditions, be better than previously understood. [25,26]. Suggestions that SST be discontinued due to scientists' apparent misunderstandings may be hasty. Of course, there may be other reasons to abandon SST, but our work shows that given the opportunity, scientists successfully deploy basic SST logic. In spite of scientists' real-life statistical behaviour often resembling a "ritual" [27], when we eliminate the ritual — no p value, or any other familiar number, was offered — they think statistically, very often arriving at the correct conclusion about the sign of the effect.

We wish to emphasize what we cannot, and do not, argue. First, we cannot argue that simply because scientists can successfully use SST logic, that they *do* in real situations, or that specific instantiations of SST, such as p values, are used well. We specifically set out to abstract the logic away from the typical situations in which scientists use the logic. This has the benefit of being helping to identify where problems might be, but the downside that generalizing the results will require further work. We also cannot address other potential arguments against SST, such as philosophical ones.

Finally, we hope to provide a fresh method and perspective on a long-standing debate in statistical cognition. Simulation-based approaches to teaching statistics have long been touted [28, 29]. Simulation-based approaches to *studying* scientists' statistical reasoning may also profitable, particularly in studying reasoning that is difficult for participants to articulate formally. If we are to reform statistical education and practice in the sciences, we should base that reform on diverse lines of evidence about scientists' reasoning. Understanding and harnessing scientists' already-existing competence in statistical reasoning is essential to developing effective methodological reforms.

8. Methods

(a) Participants

Participants were recruited via social media platforms such as Twitter and Facebook. All participants gave informed consent. Data inclusion criteria included sampling at least one shuffle report and experimental result, working in a scientific field, having at least some University education in science, and that it was their first time participating. Details are given in Supplement B.

After applying all inclusion criteria, 506 participants remained for analysis.

(b) Experimental Design and Procedure

Each participant was randomly assigned to one of eight true effect sizes (from $\delta = 0$ to $\delta = 1$) and one of two evidence powers ("wide" $q = 3$ or "narrow" $q = 7$; see "Evidence Distributions" below). The probability of being assigned $\delta = 0$ was 25%, while the remaining effect sizes were equally probable at 11%. The probability of assignment to either evidence power was 50%. Details are given in Table 1.1 in Supplement A.

After offering informed consent, participants read the cover story and instructions. During the instructions, the participant was introduced to the task through sampling random shuffle reports. After a brief recap of the instructions, participants performed the main task — sampling either random shuffles or experiments — until they made a decision about which, if either, elf group was faster. They were then asked several open-ended questions about their strategy, some informational questions (results in Supplement B) and debriefed.

Qualtrics' duration estimate indicated that the median time spent on the experiment was 21 minutes.

(c) Evidence distributions

The evidence/horizontal (x) location test statistic presented to the participant was derived from a transformed Z statistic:

$$Z \sim \mathrm{Normal}(\delta\sqrt{n/2}, 1)$$

where δ is a true effect size (randomly assigned to each participant, from 0 to 1) and n is the selected but unknown sample size (from 10 to 200 participants per group). Z then transformed to the (-1,1) space:

$$x = \mathrm{sgn}(Z)\left[1 - \left(1 - F_{\chi^2_1}\left(Z^2\right)\right)^{\frac{1}{q}}\right],\ -1 \leq x \leq 1.$$

where $F_{\chi^2_1}$ is the cumulative distribution function of a χ^2_1 random variable, and $q \in \{3, 7\}$ was randomly assigned for each participant. $x = -1$ represented the left edge of the interface, $x = 0$ the middle, and $x = 1$ the right edge. The setting of q determined how spread out the test statistic was on the display. This arbitrary transformation was done to ensure that the test statistic's distribution was unfamiliar to the participant. See Supplement A for more details, including graphical depictions of the evidence distributions.

(d) Coding of open-ended strategy questions

We determined the coding scheme and independently categorized the first 20 participant, discussing the source of disagreements. After categorizing the remaining participants, some disagreements were resolved through mutual agreement, and a discussion between the authors was had over what caused the disagreements. The remainder of the disagreements were re-coded separately, and a final round of discussion resolved the remaining disagreements. The coding of participants' responses is described in detail in Supplement B.

- **Funding**: This research was not supported by external funding.

- Compiled 2020-11-04 16:45:34 (Europe/London) under R version 4.0.3 (2020-10-10).

Ethics. This research project was evaluated by the Cardiff University School of Psychology (application number EC.18.12.11.5526G). It was found to be within the ethical guidelines for experiments with human participants. All participants gave informed consent prior to their participation.

Data Accessibility. Data and relevant code for this research work are stored in GitHub: https://github.com/richarddmorey/Morey_Hoekstra_StatCognition and have been archived within the Zenodo repository: https://doi.org/10.5281/zenodo.3877106

Authors' Contributions. RDM conceptualized and designed the study in consultation with RH. RDM analysed the data and created the materials and figures. The manuscript was written by RDM and RH.

Competing Interests. The authors declare no conflicts of interest.

References

1. Dempster AP. 1964 On the Difficulties Inherent in Fisher's Fiducial Argument. *Journal of the American Statistical Association* **59**, 56–66.
2. Greenland S. 2019 Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. *The American Statistician* **73**, 106–114. Publisher: Taylor & Francis.
3. Oakes M. 1986 *Statistical inference: A commentary for the social and behavioral sciences*. Chichester: Wiley.
4. Falk R, Greenbaum CW. 1995 Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology* **5**, 75–98.
5. Haller H, Krauss S. 2002 Misinterpretations of Significance: A Problem Students Share with Their Teachers?. *Methods of Psychological Research Online* **7**.
6. Lecoutre MP, Poitevineau J, Lecoutre B. 2003 Even statisticians are not immune to misinterpretations of Null Hypothesis Tests. *International Journal of Psychology* **38**, 37–45.
7. Kalinowski P, Fidler F, Cumming G. 2008 Overcoming the Inverse Probability Fallacy: A Comparison of Two Teaching Interventions. *Methodology* **4**, 152–158.
8. Carver R. 1978 The Case Against Statistical Significance Testing. *Harvard Educational Review* **48**, 378–399.

9. Fidler F. 2006 Should Psychology abandon p values and teach CIs instead? Evidence-based reforms in statistics education. In *Proceedings of the 7th International Conference on Teaching Statistics*.
10. The B. 2011 Significance testing - are we ready yet to abandon its use?. *Current Medical Research and Opinion* **27**, 2087–2090. PMID: 21916530.
11. Wasserstein RL, Lazar NA. 2016 The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* **70**, 129–133.
12. Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ. 2014 Robust Misinterpretation of Confidence Intervals. *Psychonomic Bulletin & Review* **21**, 1157–1164.
13. Hoekstra R, Johnson A, Kiers HA. 2012 Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement* **72**, 1039–1052.
14. Gelman A, Stern H. 2006 The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* **60**, 328–331.
15. Hoekstra R, Finch S, Kiers HAL, Johnson A. 2006 Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review* **13**, 1033–1037.
16. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. 2011 Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience* **14**, 1105–1107.
17. Weisburd D, Lum CM, Yang SM. 2003 When can we Conclude that Treatments or Programs "Don't work"?. *The Annals of the American Academy of Political and Social Science* **587**, 31–48.
18. Pashler H, Harris CR. 2012 Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science* **7**, 531–536. PMID: 26168109.
19. Kennedy-Shaffer L. 2019 Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing. *The American Statistician* **73**, 82–90. PMID: 31413381.
20. Lilienfeld SO. 2017 Psychology's Replication Crisis and the Grant Culture: Righting the Ship. *Perspectives on Psychological Science* **12**, 660–664. PMID: 28727961.
21. Ishihara S. 1972 *Tests for colour-blindness*. Tokyo: Kanehara Shuppan Co. Ltd 24 plate edition edition.
22. Venables WN, Ripley BD. 2002 *Modern Applied Statistics with S*. New York: Springer fourth edition. ISBN 0-387-95457-0.
23. Gelman A, Carlin J. 2014 Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* **9**, 641–651. PMID: 26186114.
24. Macmillan NA, Creelman CD. 2005 *Detection Theory: A user's guide*. Mahwah, N.J.: Lawrence Erlbaum Associates 2nd edition.
25. Cosmides L, Tooby J. 1992 Cognitive Adaptations for Social Exchange. In *The Adapted Mind: Evolutionary psychology and the generation of culture*, pp. 163–228. New York: Oxford University Press.
26. Gigerenzer G, Hoffrage U. 1995 How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review* **102**, 684–704.
27. Gigerenzer G, Krauss S, Vitouch O. 2004 The null ritual: What you always wanted to know about significance testing but were afraid to ask. In Kaplan D, editor, *The Sage handbook of quantitative methodology for the social sciences* , . Thousand Oaks, CA: Sage.
28. Cumming G, Thomason N, Howard A, Les J, Zangari M The StatPlay software for statistical understanding: Confidence intervals and hypothesis testing. Paper presented at the 1995 meeting of the Australian Society for Computers in Learning in Tertiary Education.
29. Rossman AJ, Chance BL. 2014 Using simulation-based inference for learning introductory statistics. *WIREs Computational Statistics* **6**, 211–221.