‘Are You Taking the *P*?’ Correcting Psychology Students’ Misconceptions About Probability Values

Abstract

Statistics are an almost unavoidable component of modern psychology and probability (or *p*) values are ubiquitous in published work. Despite this, many students hold misconceptions about what *p* values represent, some of which suggest a fundamental misunderstanding of the underlying theory. This study investigated whether it was possible to correct a specific misconception by showing participants a video that provided a detailed explanation of the correct interpretation of a *p* value. Scores on two measures of statistical knowledge suggested that the correction was ineffective. This study also found that repeating the misconception during correction had no effect on the efficacy of a correction. Challenges in correcting *p* value misconceptions are highlighted, including the key issue potentially being either an insufficiently detailed statistical education, or the complexity of the underlying theory. Implications for statistical education are discussed and Bayesian methods are proposed as a simpler alternative in psychological research and undergraduate education.

"The unity of all science consists alone in its method, not in its material"

(Pearson 1900/1892, p. 12)

Broadly speaking, scientific investigation can be split into theoretical ideas and practical research. The latter includes how experiments are planned and conducted, and how the resulting data are analysed. Despite their importance, many undergraduate students do not enjoy learning these practical elements (Murtonen, 2005; Murtonen & Lehtinen, 2003), preferring instead to learn the theoretical ideas that underpin experiments (Vittengl et al., 2004).

Learning statistical methods is often regarded as the least enjoyable component of undergraduate psychology programmes (Addison, Stowell, & Reab, 2015), and students even report statistics courses to be anxiety-inducing (Hanna, Shevlin, & Dempster, 2008; Macher, Papousek, Ruggeri, & Paechter, 2015). Unsurprisingly, this leaves students with a weak grasp of statistical concepts (Macher et al., 2015). This a major concern given that the majority of modern psychological investigations include statistics. It is therefore essential that students graduate with a solid understanding of how to interpret them.

A starter in statistics

By far the most prevalent school of statistical analysis in psychology is frequentist, taught in almost all universities and reported in almost all published papers (Hubbard & Ryan, 2000). Frequentist methods were developed in the early 20th century by notable mathematicians such as Egon Pearson, Jerzy Neyman and Ronald Fisher, in an attempt to provide previously anecdotal sciences with methodological rigour (Perezgonzalez, 2015).

The majority of statistical tests in psychology are null hypothesis significance tests. The result of which is a probability (or *p*) value, representing the probability of observing a result given that the null hypothesis is true (Hubbard & Armstrong, 2006). *P* values are the most prevalent statistic in published work, with 94% of published psychological papers reporting at least one (Hubbard & Ryan, 2000).

The inferential utility of *p* values is dependent on many factors including sample size and effect size (Ziliak & McCloskey, 2008). However, results are often judged to have merit solely on the statistical significance of a *p* value, and many authors fail to report other crucial results (Fritz, Scherndl, & Kuhberger, 2013; Kühberger, Fritz, Lermer, & Scherndl, 2015). This is dangerous practice, as *p* values by themselves are a relatively uninformative measure (Armstrong, 2007; Kline, 2013).

Despite their prevalence, *p* values have always been marked by issues, both theoretical and practical (Benjamin et al., 2018; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008; Wagenmakers, Marsman, et al., 2018). One of these practical issues is the prevalence of misconceptions (false ideas or beliefs) regarding what *p* values represent and how they should be interpreted (Cohen, 1994; Kirk, 2001).

Misconception prevalence

Replicated findings show that 90% of psychologists hold at least one misconception regarding what *p* values represent (Haller & Krauss, 2002). Even statistics textbooks and 80% of statistics teachers have been found to hold misconceptions, making it likely that they will be passed on to students (Badenes-Ribera & Navarro, 2017; Gigerenzer, Krauss, & Vitouch, 2004; Gliner, Morgan, Leech, & Harmon, 2001; Haller & Krauss, 2002; Lecoutre, Poitevineau, & Lecoutre, 2003). Misconceptions are even more problematic when they make it into published work. Findings suggest that up to 18% of reputable journal articles contain incorrectly reported statistics (Bakker & Wicherts, 2011). Incorrect statistics lead to inappropriate conclusions which seriously damage the integrity of the literature and are a key contributor to the field's current replication crisis (Aarts et al., 2015).

Misconceptions are most prevalent in undergraduate students. One study found that all 44 students tested held at least one of six misconceptions, with some students holding multiple incorrect views as to what a *p* value represents (Haller & Krauss, 2002). *P* values are by no means the only component of hypothesis testing subject to misconceptions (Chance, del Mas, & Garfield, 2004; Lipson, 2002; Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). However, given the prevalence and influence of *p* values, this is an issue which warrants attention.

Misconception content

Many different misconceptions regarding *p* values exist (Goodman, 2008). The following two are understood to be among the most common (Kline, 2013), with *p* values being (incorrectly) interpreted as the probability that:

1) The results are due to chance;
2) The research hypothesis is true.

Both of these interpretations are incorrect. A *p* value represents the probability of observing the data given the null hypothesis is true. The first of these misconceptions is regarded as being the most prevalent (Carver, 1978; Kline, 2013), however the second misconception is likely to be the most damaging.

Paradoxically, frequentist hypothesis tests tell us nothing about the probability of hypotheses. Instead, they report the probability of *observing the data given the hypothesis*, a subtle but important difference (Wagenmakers et al., 2008). Thinking that a *p* value represents the probability of a hypothesis being true (or false) is a fundamental misunderstanding of the theory that underlies the method.

Up to 59% of undergraduate students believe this misconception (Haller & Krauss, 2002). Given the prevalence of *p* values in psychological research, it is concerning that so many students demonstrate such a profound misunderstanding of what they represent. To date, relatively few studies have attempted to correct *p* value misconceptions (Khazanov & Prado, 2010; Krauss & Wassner, 2002). Fortunately, a wealth of literature on correcting misconceptions already exists.

Correcting misconceptions

Part of the challenge in communicating a correct interpretation of a *p* value is its complexity, with understanding requiring the integration of many abstract ideas (Sotos et al., 2007). Humans have a bias towards believing simple explanations, increasing the risk of misconceptions when communicating complicated ideas (Chater & Vitanyi, 2003; Lombrozo, 2007; Pacer & Lombrozo, 2017). However, it is possible to overcome this bias by providing participants with a detailed understanding of why misconceptions are incorrect (Kowalski & Taylor, 2009; Weisman & Markman, 2017).

Much of the existing literature on correcting misconceptions looks to change only a single fact or an incorrect headline, limiting its use in the current context (Chan, Jones, Jamieson, & Albarracin, 2017). The literature does, however, highlight three important considerations in devising and delivering an effective correction: narrative coherence, backfire effects, and repetition.

Narrative coherence

Participants likely hold internal narratives which structure their knowledge and therefore contain any misconceptions (Johnson & Seifert, 1994). To enable a complex idea to replace a misconception, the information needs to fit into a coherent narrative (Johnson & Seifert, 1994; Johnson-Laird, 2012; Schwarz, Sanna, Skurnik, & Yoon, 2007). If a correction changes facts but does not provide a coherent narrative, participants are more likely to rely on their previous misconception, *even if they know it to be* *false* (Gerrie, Belcher, & Garry, 2006; Johnson & Seifert, 1994).

Providing an alternate narrative is more effective than disputing facts (Tenney, Cleary, & Spellman, 2009), and allowing participants to reason through cognitive conflicts and see why a misconception cannot be correct is also beneficial (Khazanov & Prado, 2010; Rapp & Kendeou, 2007; Seifert, 2002). Given that the correct interpretation of a *p* value is the only true interpretation, there *is* a coherent narrative behind it. The challenge is communicating this simply and with sufficient detail.

Backfire effects

Backfire effects arise when an attempted correction paradoxically *increases* the strength of a misconception in a participant's mind (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Peter & Koch, 2016). Backfire effects would mark an attempted correction as a grievous failure and it is therefore essential that they are avoided.

Theories as to why backfire effects exist suggest that entrenched ideological beliefs cause participants to think that views that require challenging must be credible (Nyhan & Reifler, 2010). Such concerns are unlikely to affect the issue at hand as students – whilst not liking statistics – probably do not have strong ideological beliefs about what *p* values represent. Of more relevance, backfire effects can arise when corrections are too complicated for participants to understand. This highlights the need for a coherent correction (Chan et al., 2017). Indeed, there is evidence to show that corrections can be effective – and backfire effects minimal – when information is presented clearly (Wood & Porter, 2019).

With these two issues in mind, the correction used in this study is based on providing a detailed and coherent narrative of what a *p* value represents. This is similar to how statistics are taught in the first instance, meaning that if the correction is successful, implementation in education is realistic.

Repetition

A final consideration is whether or not the misconception is repeated during the correction. Many studies find that repeating the misconception (even whilst correcting it) strengthens belief in it, as the information seems more familiar (Eakin, Schreiber, & Sergent-Marshall, 2003; Ecker, Lewandowsky, Swire, & Chang, 2011; Lewandowsky et al., 2012).

However, avoiding repetition of the misconception may lead participants to become confused as to what the correction relates to (Chan et al., 2017). Therefore, it may be necessary to repeat the misconception when providing a detailed correction. There is also evidence to show that the negative effects of repetition can be avoided if the misconception is refuted as it is mentioned (Cook & Lewandowsky, 2011).

The present study investigates the effect of repetition by using two experimental conditions: one which repeats and immediately refutes the misconception, and one that does not repeat the misconception at all.

The present study

This study investigated whether it was possible to correct a misconception by showing participants a detailed video explanation of the correct interpretation of a *p* value. The video showed a statistics teacher and accompanying slides and was chosen to be analogous to classroom learning, therefore more representative of how students learn statistics. In addition, video supplements are increasingly popular in education (Gedera & Zalipour, 2018; McGarr, 2009), and there is evidence to suggest that video is a more effective medium for communicating information than writing (Wilson et al., 2012; Wilson et al., 2010).

The present study aims to correct the misconception that a *p* value represents the probability of the research hypothesis being true. This misconception is especially prevalent in students and demonstrates a crucial misunderstanding of the underlying statistical theory. The correction aims to provide a detailed explanation of what a *p* value represents and why, enabling participants to understand why the misconception is incorrect.

H1 – Participants shown a corrective video will score higher on tests of statistical knowledge than those not shown a corrective video.

I also investigated the effect of repeating the misconception on the efficacy of a correction. Given that the correction is detailed, I did not expect backfire effects to be present. However, in line with the literature, I expected that the correction with repetition would be less effective than the correction without repetition.

H2 – Participants shown a corrective video that repeats the misconception will score lower on tests of statistical knowledge than participants shown a corrective video that does not repeat the misconception.

Method

Design

This is a between-subjects design with one independent variable (condition: Control, No repetition, Repetition), and two dependent variables that measure statistical knowledge. This study was implemented as part of a larger study which included two additional conditions and two additional dependent variables. Each participant completed one condition and all four dependent measures.

Participants

A frequentist *a priori* power analysis was conducted using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), based on the results of a meta-analysis which reported an effect size of between *d* = 1.14 – 1.33 for the correction of misinformation (Chan et al., 2017). Given the sheer magnitude of this effect and the tendency for reported effects to be inflated (Aarts et al., 2015; Bakker, van Dijk, & Wicherts, 2012), a more conservative estimate of effect size (*d* = 0.6) was used. With alpha level .05 and power of .8, a target sample size of 36 participants per condition was obtained, for a total of 108.

Due to time limitations and in accordance with the stopping rules specified in the pre-registration (https://osf.io/7gex9/), a total of 88 participants completed the study, all of whom were psychology students at The University of Bristol. Due to the lower participant numbers, a sensitivity power analysis was conducted using G*Power to determine the minimum detectable effect (MDE) that could be observed for each of the comparisons. This yielded an MDE of *d* = 0.56 for the comparison between control and experimental conditions, and *d* = 0.67 for the comparison between the two experimental conditions.

Those in the first and second year of their undergraduate degree completed the experiment in exchange for course credit. Others were not reimbursed for their time. Full demographic information can be seen in Table 1.

Table 1.
Demographic information for each condition

			Age		Degree Year (% breakdown)			
Condition	*N*	% Female	*M*	*SD*	First	Second	Third	Masters
Control	31	77	20.23	1.77	35.48	45.16	16.13	3.23
No repetition	28	92	20.43	2.22	35.71	35.71	21.43	7.14
Repetition	29	82	21.34	6.37	27.59	44.83	27.59	0.00
Total	88	85	20.70	3.99	32.95	42.05	21.59	3.41

Materials and Measures

The experiment centred around a fictional news report which described a fictional study conducted in the Netherlands. This tested the effect of reintroducing wild animal species on the number of flood days per year in 24 rural areas. The full article can be seen in the online supplement (https://osf.io/7gex9/). The report was designed to simply outline the structure of a frequentist independent samples *t*-test. The report highlights that the researchers' investigation yields a result significant at the $p<.05$ level. Crucially, the researchers (incorrectly) conclude that a *p* value corresponds to the probability of the research hypothesis being true.

Independent variable

In order to correct the misconception, two short videos were produced. These involved a well-known statistics teacher at The University delivering a script whilst a simple graphic beside him illustrated how a statistical result derived from a sample corresponds to the distribution of differences in a population (see Figure 1). The statistics teacher was used in order to provide the correction with credibility.

The script for these videos was written to correct the misinterpretation of the *p* value presented in the text. The two scripts differed only in whether or not they repeated the misconception from the text whilst correcting it. The No repetition script did not repeat the misconception whilst the Repetition script repeated it four times (see online supplement).

Figure 1. Screenshots from the corrective videos.

Dependent variables

The first dependent variable was a single open question, asking 'How should a *p* value be interpreted?'. This was designed to test participants' knowledge of *p* values outside of the strict context encountered in the experiment. There was no minimum or maximum word count.

The answers to this question were analysed by three raters (including the author) and scored from 0 – 3. To make the analysis as transparent and objective as possible, a keyword marking system was used whereby answers were awarded one mark for including each of the following three phrases (or variations of them): 'the probability of', 'observing these results', 'given the null hypothesis is true'. The three raters independently rated answers and then discussed any inconsistencies to reach a single set.

As a separate analysis, answers were also classified by the raters into three open question groups depending on whether answers were 1) unambiguously correct, 2) repeated the misconception from the article, or 3) either evidenced a different misconception or were incorrect.

The second dependent variable was participants' score on six closed ended sentence completion questions. Participants saw the sentence stem from the initial article and clicked to select whether each presented ending represented a correct (2/6) or incorrect (4/6) interpretation of the researchers' results. These were adapted from a previous study, and designed to specifically test four misconceptions (see Table 2) (Haller & Krauss, 2002). Each participant received a score out of six corresponding to how many questions they answered correctly.

Table 2.
The closed questions that formed the second dependent variable and the misconception each was designed to test

Sentence ending	Misconception tested
There is a less than 5% likelihood that rewilding had no effect on flood days.	*P* is the probability of the null hypothesis being true.
These results were at least 95% likely to have occurred due to the rewilding project.	*P* is the probability of the alternative hypothesis being true.
There is a less than 5% likelihood that accepting the effects of the rewilding project as true is the wrong decision.	*P* is the probability of making a type I error.
If the rewilding experiment could be repeated 100 times, the result would be significant at the $p<.05$ level 95 out of the 100 times.	*P* is the probability of the same result being obtained through replication.
There was a less than 5% likelihood of these results having occurred even if the rewilding project had no effect.	Correct interpretation.
There was a less than 5% likelihood that something other than the rewilding project caused the reduction in the number of flood days.	Correct interpretation.

Note. Sentence prompt: 'The researchers found that the rewilded areas experienced fewer flood days, a finding that was significant at the $p<.05$ level. This means that _______.'

(Adapted from Haller & Krauss, 2002).

Procedure

The experiment was delivered through the Qualtrics online platform (Qualtrics, 2019). Of the 88 participants, 79 came into the lab to do the experiment and 9 (three from each condition) took the experiment remotely on either a desktop or laptop in a push to increase sample size. Participants were instructed to wear the supplied headphones even though they may not need them. Participants were randomised into a condition by Qualtrics and instructions were provided through the software.

After proving demographic information and consent, participants read the article which contained the misconception. Participants in the two experimental conditions then saw a screen introducing the video and watched the video. The videos were 2:25 and 2:46 minutes long, for the No repetition and Repetition conditions respectively. Participants were free to rewind and skip through the video as they desired and the page auto-advanced after 4:00 minutes. All participants then answered several arithmetic questions as a short distractor task before the questions. These were not assessed and the page ensured participants spent between 1:30 and 2:00 minutes on the task.

Participants responded to four sets of questions: two open questions, and two banks of six closed questions. The open questions were shown first. The first asked them to complete the final sentence of the article so it represented a correct interpretation of the researchers' results. The second question, as outlined in the dependent variables section, asked how a *p* value should be interpreted.

Within the two banks of closed questions, each of the six possible answers appeared on screen one at a time and in a randomised order. The first bank is shown in Table 2. The second bank of questions presented sentences and asked if each was a correct or incorrect interpretation of what a *p* value represents. Only the data from the second of the open questions and the first bank of the closed questions are reported here. The data from the other questions are reported in a separate study. The order of the question sets was not counterbalanced as this would have given some participants an advantage. Participants were then debriefed. The experiment took between 7.5 and 25 minutes, depending on condition. The study was approved by The University's Research Ethics Committee (code: 80902).

Data analysis plan

As part of the effort to avoid misconceptions, the data here are analysed using Bayesian methods. A parallel, frequentist analysis is presented in the online supplement for comparison. The two analyses suggest similar conclusions.

The key hypothesis tests require a comparison of independent participant groups on two dependent measures. This is split into four Bayesian independent samples comparisons (analogous to independent samples *t*-tests), two for each dependent variable. The comparisons are the same for both dependent variables. The first compares the scores of the control condition to the average scores of both experimental conditions to test whether there is an effect of correction. The second compares the scores of the two experimental conditions to test whether there is an effect of repetition. For all four of these comparisons, a Bayes factor is presented. In addition, effect sizes and 95% credible intervals are provided. A Bayesian contingency table was also created as an exploratory measure to investigate how each condition affected the prevalence of the original misconception after correction.

It is worth noting that Bayesian comparisons do not need to correct for inflated error rates arising from multiple comparisons in the same way that frequentist methods do as Type I error does not exist in Bayesian analysis (Gelman, Hill, & Yajima, 2012).

Analysis

Two comparisons were conducted on open question score. The first was between the scores for the Control condition and the average of the two experimental conditions; the second comparison was between the two experimental conditions. The same comparisons were also conducted on closed question score, the second dependent variable.

All Bayesian comparisons used the default Cauchy prior width (r = .707), a conservative prior that slightly favours the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Love, et al., 2018). As tests were confirmatory, a directional hypothesis was used in all tests. It was hypothesised that on both dependent variables, the average score for the experimental conditions would be higher than the average for the Control condition, and that the average for the No repetition condition would be higher than that of the Repetition condition.

The result of a Bayesian analysis is a Bayes factor (BF) which represents the likelihood of a hypothesis given the observed data. All Bayes factors are reported as BF10s, representing the strength of evidence in favour of the research hypothesis. In line with the classification scale developed by Wagenmakers and colleagues (2018), BFs between .33 and 3 are seen as non-diagnostic and those between 3 and 10 represent moderate evidence in favour of the research hypothesis. BFs 10 to 30, 30 to 100, and >100 represent strong, very strong, and extreme evidence in favour of the research hypothesis respectively. In contrast, BFs .33 to .10, .10 to .03, .03 to .01, and <.01 represent moderate, strong, very strong, and extreme evidence in favour of the null hypothesis respectively. The Bayesian analysis was conducted using JASP 0.9.2 (JASP Team, 2019).

Table 3.
Means, Bayes factors (BF), effect sizes, and credible intervals for the four comparisons

	Descriptive statistics			Comparisons		
Condition	*N*	*M*	*SD*	Mean difference	BF	∂ [95% Credible Interval]
				Open Question Scores		
Control	31	0.71	0.86			
Experimental	57	1.11	1.15	0.40	1.47	0.33 [-0.07, 0.76]
No repetition	28	0.96	1.07	-0.28	.016	-0.20 [-0.69, 0.27]
Repetition	29	1.24	1.22			
				Closed Question Scores		
Control	31	2.71	1.22			
Experimental	57	3.11	1.26	0.40	1.05	0.27 [-0.14, 0.71]
No repetition	28	2.96	1.32	-0.28	0.16	-0.18 [-0.67, 0.30]
Repetition	29	3.24	1.22			

Note. The BFs are placed in the row of the condition hypothesised to have the higher score.

Note. The mean differences for the two dependent variables are *exactly* the same. I am aware that this is more likely due to researcher error than chance (Abelson, 1995), and I have checked the data entry and analysis to ensure its legitimacy. Satisfied with the integrity of the outcome, I have made the raw data available at https://osf.io/7gex9/ for inspection should the reader wish to do the same.

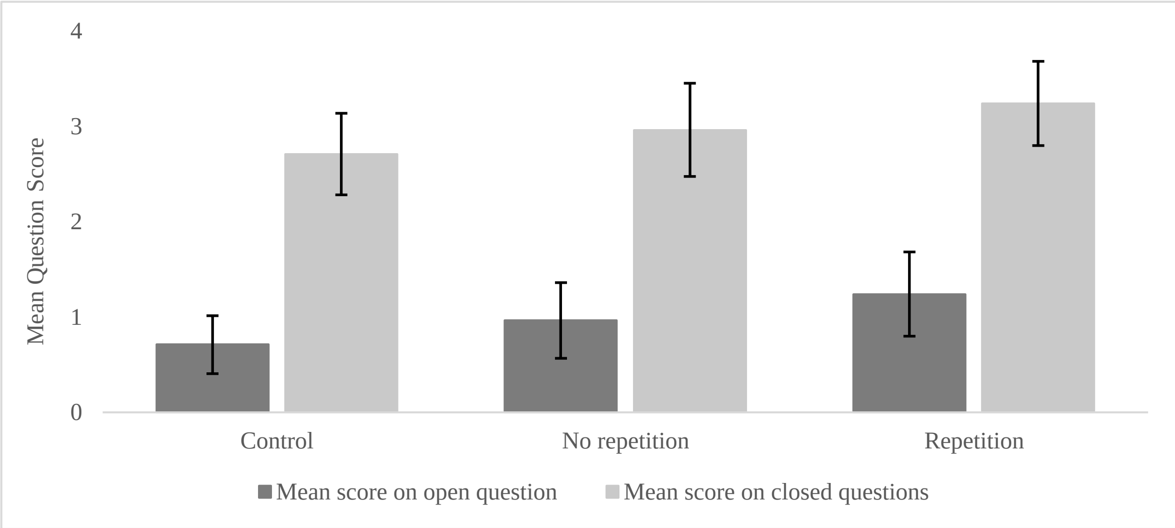

Figure 2. Mean scores and credible intervals for open and closed questions across conditions.

Effect size ∂ is the standardised difference between two independent means (Cohen's *d* for the population) and the 95% credible interval represents a range we can be 95% confident that the true value of ∂ falls within. In calculating these, a non-directional hypothesis was used in order to allow the credible interval to span both sides of 0.

Bayes factors, effect sizes and credible intervals for each of the four pairwise comparisons are reported in full in Table 3. Mean scores on both dependent variables across the three conditions are presented in Figure 2. The prior and posterior distributions for all four comparisons can be seen in Figure 3, showing how estimates of the true population effect size were informed by both the default prior distribution used and the observed data.

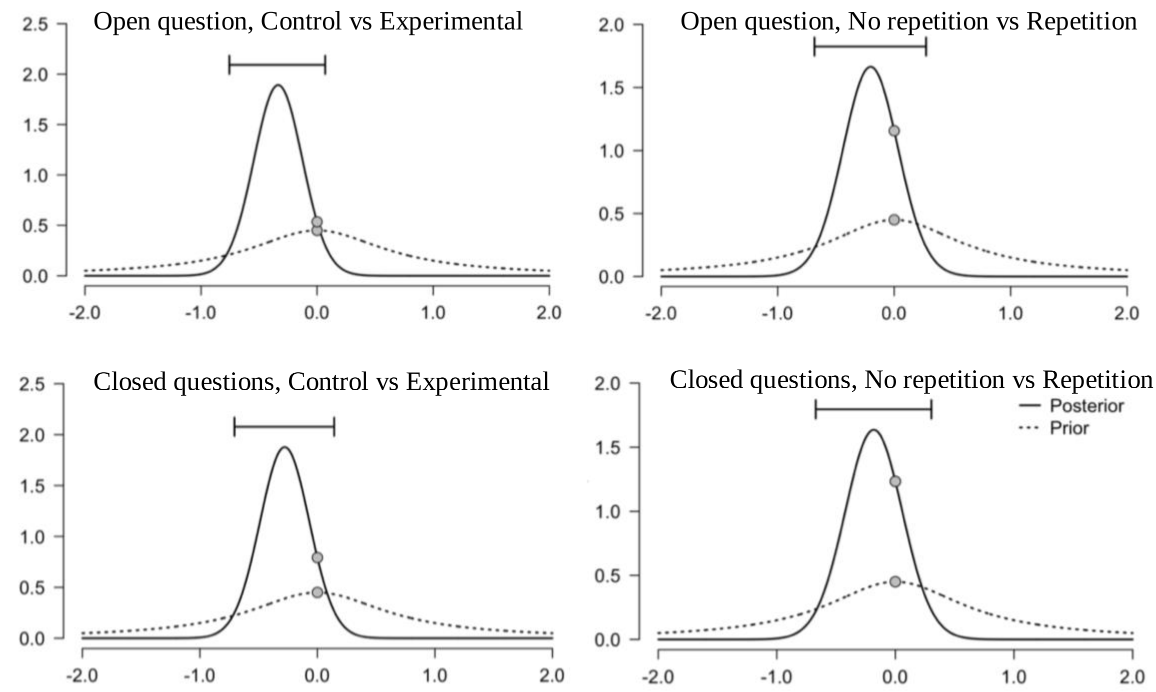

Figure 3. Showing the prior and posterior distributions for each of the four pairwise comparisons, where effect size ∂ is shown on the X-axis and density is shown on the Y-axis.

Results

In sum, the comparisons between the control condition and the mean of both experimental conditions were non-diagnostic for both dependent variables. This suggests insufficient evidence to conclude that watching a corrective video does or does not improve participants' performance on tests of *p* value knowledge.

For both dependent variables, there was moderate evidence in favour of the null hypothesis for the comparisons between the No repetition and Repetition conditions. A consultation of mean scores in Figure 2 suggests that the hypothesised effect may exist in the opposite direction, however the choice of directional hypotheses in this study prevents further comment. This suggests that repetition did not have an effect on the efficacy of misconception correction. For all comparisons, effect sizes were small and credible intervals were wide, supporting the non-diagnostic nature of the observed Bayes factors.

To assess the impact of condition on open question group, a Bayesian contingency table was constructed (equivalent to a frequentist chi-squared test). Overwhelmingly, participants in all conditions provided incorrect answers to the open question (see Figure 4).

The analysis yielded a Bayes factor of BF01 = 12.03, meaning the null hypothesis (that there is no difference between groups) was 12 times more likely than the alternative hypothesis.

Discussion

This study aimed to investigate the effect of corrective videos on the prevalence of a misconception about what a *p* value represents. This study found no diagnostic evidence to suggest that participants shown a corrective video demonstrated fewer misconceptions than those not shown a corrective video. This study also investigated the effect of repeating the misconception and found moderate evidence to suggest that the efficacy of a corrective video was not affected by repetition of the misconception.

The correction

Perhaps the most likely explanation for the failure to correct the misconception is the quality of the correction itself. Narrative coherence was the key objective in devising the correction and one which I believe to have been met. However, embedding the correction into a narrative makes it longer and it would be interesting to investigate whether the length of a correction affects efficacy.

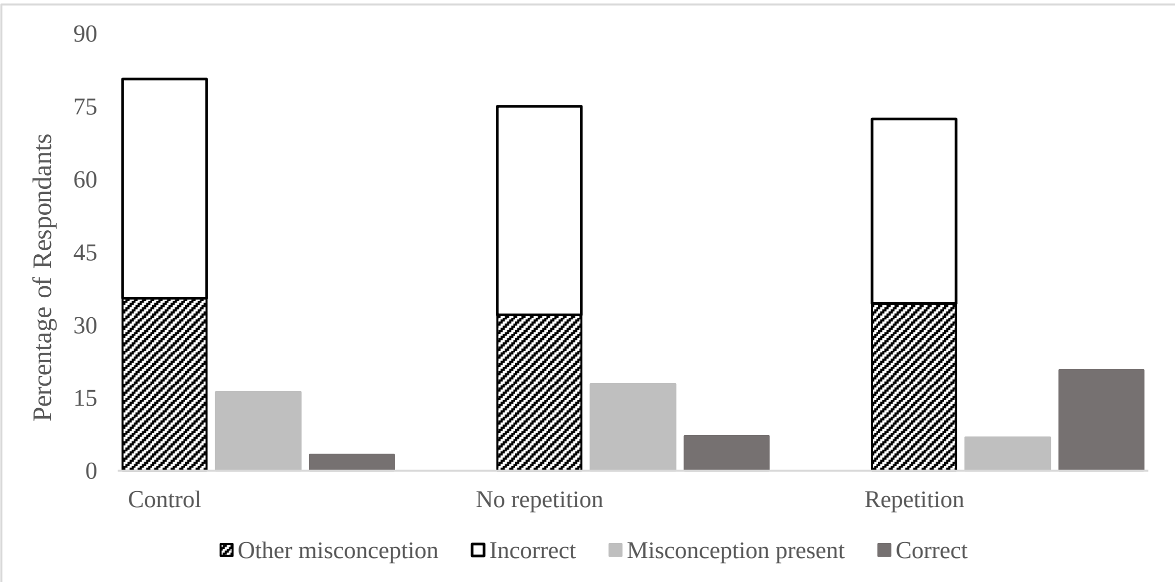

Figure 4. Percentage breakdown of open question group by condition.

The correction was devised over a period of several weeks and involved lengthy discussion between six psychology students and a statistics teacher. The online supplement (https://osf.io/7gex9/) contains the correction in full and I welcome any comments on how it can be improved, without detracting from its statistical accuracy or brevity. Interestingly, trying to communicate this idea in a simple way was a real challenge, and may suggest that the inherent complexity of the subject matter is why misconceptions are so prevalent in textbooks (Gliner et al., 2001).

Another point to note is the number of participants who showed some form of misconception other than the one explicitly tested for (34.04%) (see Figure 4). This was relatively stable across conditions and may suggest that any correction needs to be tailored to the specific misconception held by students.

One potential limitation of this study is the use of a video correction. Recent results tentatively suggest that a written correction may be more effective (Peachey, 2019), and future work comparing the efficacy of different correction methods could better inform interventions. It is also worth noting that there was a difference between how the misconception and the correction were presented, in writing and as a video respectively. Future work could investigation whether the efficacy of corrections is higher when they are presented in the same medium as the misconception.

A definite limitation is the study's low statistical power. The observed effect sizes were smaller than the minimum detectable effect gained from the sensitivity power analysis. This means that whilst the true population effect could be of the magnitude reported, the study is insufficiently powered to detect an effect this small. This could be remedied by testing more participants. If future work used Bayesian methods, the present results could be incorporated as a prior belief, providing a closer estimate of the true population effect (Wagenmakers, Marsman, et al., 2018).

Repetition

The repetition of the misconception during correction was hypothesised to make the correction less effective. The results of this study did not support this hypothesis, with the relationship appearing to exist in the opposite direction (see Figure 2). This suggests that repetition may not be a concern for detailed corrections. Such a finding is useful in informing interventions and is in line with existing evidence (Cook & Lewandowsky, 2011). However, it is possible that the effects of repetition were too small to be observed in this study simply because the correction worked on so few participants. It is therefore difficult to conclude on whether the efficacy of a correction is affected by repetition.

To my knowledge, this is the first study that seeks to correct misconceptions students hold about *p* values. Given the novelty of the research question and the low power of this study, these findings should not be taken to conclude that *p* value misconceptions cannot be corrected. Instead, this study can offer tentative suggestions about the nature of the issue and raise interesting points for future research.

Correction vs education

One important question is whether we should consider this an issue of misconception correction, or one of statistical education. As noted in the introduction, many important factors in correcting misconceptions are not relevant in the current context. Students don't hold strong ideological beliefs surrounding *p* values nor are they under any obvious social pressure to continue believing misconceptions. In addition, undergraduates tend to have a weak grasp of statistics and scientific research methods in general, with most students seeing statistical analysis as process-driven (Rothman, 2014). In contrast, statistically adept scientists need detailed knowledge of a range of useable methods, and good judgement concerning when to apply them (Krueger & Heck, 2019).

Therefore the issue may be that scientific education does not teach enough of the statistical theory underlying the interpretation of *p* values, leaving students susceptible to fill their knowledge gaps with misconceptions. If *p* value misconceptions are stubborn through students' ignorance of the relevant theory, the challenge becomes teaching statistics simply and in enough detail that misconceptions cannot take root.

Education as antidote

Students need to be able to recognise and employ good scientific practices from an early stage. This is enabled through making scientifically rigorous methodologies accessible and easily understandable. This study, and the literature on statistical education contain several suggestions for how this can be achieved.

One of the most well supported findings is to teach statistics grounded in real research, enabling students to feel the subject is tangible rather than theoretical (Williams, McCutcheon, Fava, & Aruguete, 2017). Such a method was used in the present study, with both the theory and the misconception tied to a practical example. Making the subject more easily understandable can also counter students' dislike of statistics, which may encourage students to spend more time grappling with difficult concepts (Huynh & Baglin, 2017).

Another important consideration is the number of statistics teachers who hold misconceptions (Haller & Krauss, 2002). Training teachers to spot and correct their own misconceptions could prevent them being passed to students, and so is an important step in addressing the issue (Khazanov & Prado, 2010). However, it is not immediately clear how this could be achieved as correcting misconceptions is the very challenge this study attempts to address.

Problematically, *p* value misconceptions currently persist despite many statistics courses already being grounded in research, and the plethora of resources which offer in depth explanations of topics (Earley, 2014; Greenland et al., 2016). It is therefore worth considering that the main issue may not be a lack of education, but the complexity of *p* values themselves.

A broken paradigm

P values are complicated to understand. The fact that this and similar papers exist is testament to that. When misconceptions persist into published work, the issue is far more serious and threatens to undermine the content and reputation of psychology (Armstrong, 2007; Hubbard & Armstrong, 2006; Meehl, 1967). Researchers and journal editors should not be making such errors, but some fault may lie with the methods themselves. Even Fisher struggled to explain the inferential meaning of *p* values despite having practically created them (Goodman, 2008).

There are many ways to reach a *p* value, and researchers have to make many decisions throughout an analysis (Gelman & Loken, 2013). This means that analyses quickly lose transparency and become difficult for authors to communicate and for readers to understand (Simmons, Nelson, & Simonsohn, 2011). Some of these issues can be avoided by calculating effect sizes and confidence intervals instead of *p* values, or altering the language and processes we use in analyses (Cumming, 2008, 2014; Hurlbert & Lombardi, 2009). But such solutions do not address the root cause of the problem, namely the complexity of frequentist statistics.

Simplifying statistics could increase the transparency of research, reduce the number of incorrectly reported findings, and reduce the susceptibility of psychologists to misconceptions. As this paper has already alluded to, another school of statistics exists, one that seems to fill all of these criteria: Bayesian statistics.

Moving forwards

Although not known to most, Bayesian statistics have a longer history in scientific investigation than frequentist methods (McGrayne, 2011). Where frequentist statistics consider each analysis as one of an infinite number of notional (i.e. imaginary) replications, Bayesian methods update the probabilities of hypotheses as more data becomes available, making them more dynamic and useable. The underlying theory is simple: 'given the data observed, how likely is this hypothesis to be true?' The theoretical benefits of Bayesian statistics are numerous (Wagenmakers, Marsman, et al., 2018), but perhaps the most relevant advantage is their simplicity.

This study considers the misconception that a *p* value is the probability of the research hypothesis being true, which is precisely the *correct* way to interpret a Bayes factor (Kline, 2013). Bayes factors are intuitive and represent what many researchers want *p* values to be (Kruschke & Liddell, 2018). This intuitiveness is advantageous in avoiding misconceptions and fits well with the cognitive biases humans have towards simple explanations (Chater & Vitanyi, 2003; Lombrozo, 2007). It is difficult to see why so much time should be devoted to frequentist statistics, especially when learning so often comes at the expense of a basic awareness of Bayesian methods (Kline, 2013).

Relatively little research investigates misconceptions held about Bayesian statistics. Although I would expect them to be far less prevalent, such misconceptions could be easily investigated by applying the methods used in this and related papers. This would be a helpful step in informing reforms to statistical education.

Frequentist and Bayesian methods can also be successfully taught in concert, emphasising the differences between them (Greenland & Poole, 2013). Indeed, this has recently been incorporated into the British Psychological Society's teaching guidance on undergraduate research methods courses (British Psychological Society, 2017). This enables students to see that there are options in statistical analysis, creating well-informed scientists.

Bayesian methods are already becoming more popular in published work (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017; Wagenmakers, 2007). As this shift occurs, it makes sense for undergraduate education to do the same, so that students can understand and eventually contribute to the literature.

Conclusion

Misconceptions regarding *p* values have been shown to be prevalent in psychologists of all levels, which is concerning given their prevalence in published work. This study found that corrective videos did not reduce students' misconceptions about what *p* values represent, and that repeating the misconception had no effect on the efficacy of the correction.

It is not immediately apparent how a more effective *correction* could be devised. Instead, the issue may lie with a scientific education that leaves students without sufficient statistical knowledge to dispute misconceptions. Reforming education is a drastic but necessary step to address these misconceptions. The literature provides some suggestions as to how this could be done, however the main obstacle may be the complexity of frequentist statistics themselves.

An alternative statistical method both for general use and undergraduate education is Bayesian; the result of a Bayesian analysis being what the majority of students think a *p* value is. This intuitiveness combined with the theoretical benefits and the ease of communicating such methods makes it difficult to argue against their use. Their inclusion in statistics classrooms has the potential to benefit students, teachers, and the discipline more broadly. Such a prospect is exciting to any true scientists – those who prize the rigorous methods of their investigations above the rewards that come from the content of their findings. To paraphrase Dickens (1932/1859, p. 384), it is a far, far better test that I go to than I have ever been shown.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., . . . Open Sci, C. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716

Abelson, R. (1995). On suspecting fishiness. *Statistics as principled argument*, 78-88.

Addison, W. E., Stowell, J. R., & Reab, M. D. (2015). Attributes of introductory psychology and statistics teachers: Findings from comments on RateMyProfessors. com. *Scholarship of Teaching and Learning in Psychology, 1*(3), 229.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23*(2), 321-327. doi:10.1016/j.ijforecast.2007.03.004

Badenes-Ribera, L., & Navarro, M. D. F. (2017). Fallacies about p-Value Shared by Professors and University Students. *Universitas Psychologica, 16*(3). doi:10.11144/Javeriana.upsy16-3.fvcp

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666-678. doi:10.3758/s13428-011-0089-5

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10. doi:10.1038/s41562-017-0189-z

British Psychological Society (2017). *Supplementary guidance for research and research methods on Society accredited undergraduate and conversion programmes*. Retrieved from https://www.bps.org.uk.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399.

Chan, M. P. S., Jones, C. R., Jamieson, K. H., & Albarracin, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science, 28*(11), 1531-1546. doi:10.1177/0956797617714579

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distribitions. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323): Springer.

Chater, N., & Vitanyi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*(1), 19-22. doi:10.1016/s1364-6613(02)00005-0

Cohen, J. (1994). THE EARTH IS ROUND (P-LESS-THAN.05). *American Psychologist, 49*(12), 997-1003. doi:10.1037//0003-066x.49.12.997

Cook, J., & Lewandowsky, S. (2011). *The Debunking Handbook*: St. Lucia, Australia: University of Queensland.

Cumming, G. (2008). Replication and p Intervals p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science, 3*(4), 286-300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7-29. doi:10.1177/0956797613504966

Dickens, C. (1932). *A tale of two cities*. London: Thomas Nelson and Sons. (Original work published 1859)

Eakin, D. K., Schreiber, T. A., & Sergent-Marshall, S. (2003). Misinformation effects in eyewitness memory: The presence and absence of memory impairment as a function of warning and misinformation accessibility. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*(5), 813-825. doi:10.1037/0278-7393.29.5.813

Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teaching in Higher Education, 19*(3), 242-253. doi:10.1080/13562517.2013.860105

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review, 18*(3), 570-578. doi:10.3758/s13423-011-0065-1

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/bf03193146

Fritz, A., Scherndl, T., & Kuhberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology, 23*(1), 98-122. doi:10.1177/0959354312436870

Gedera, D., & Zalipour, A. (2018). *Use of interactive video for teaching and learning.* Paper presented at the Australasian Society for Computers in Learning in Tertiary Education, ASCILITE).

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211. doi:10.1080/19345747.2011.618213

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Gerrie, M. P., Belcher, L. E., & Garry, M. (2006). 'Mind the gap': False memories for missing aspects of an event. *Applied Cognitive Psychology, 20*(5), 689-696. doi:10.1002/acp.1221

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). *The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask.* Paper presented at the Handbook on Quantitative Methods in the Social Sciences. Sage, Thousand Oaks, CA.

Gliner, J. A., Morgan, G. A., Leech, N. L., & Harmon, R. J. (2001). Problems with null hypothesis significance testing. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(2), 250-252. doi:10.1097/00004583-200102000-00021

Goodman, S. (2008). A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology, 45*(3), 135-140. doi:10.1053/j.seminhematol.2008.04.003

Greenland, S., & Poole, C. (2013). Living with P Values Resurrecting a Bayesian Perspective on Frequentist Statistics. *Epidemiology, 24*(1), 62-68. doi:10.1097/EDE.0b013e3182785741

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. doi:10.1007/s10654-016-0149-3

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research, 7*(1), 1-20.

Hanna, D., Shevlin, M., & Dempster, M. (2008). The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences, 45*(1), 68-74. doi:10.1016/j.paid.2008.02.021

Hubbard, R., & Armstrong, J. S. (2006). Why we don't really know what

statistical significance means: Implications for educators. *Journal of Marketing Education, 28*(2), 114-120.
Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681. doi:10.1177/00131640021970808
Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*(5), 311-349. doi:10.5735/086.046.0501
Huynh, M., & Baglin, J. (2017). Teaching statistics through data investigations in Australian secondary schools: An island-based pilot project. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International), 25*(1).
JASP Team (2019). JASP (Version 0.9.2) [Computer software]. Retrieved from https://jasp-stats.org.
Johnson, H. M., & Seifert, C. M. (1994). SOURCES OF THE CONTINUED INFLUENCE EFFECT - WHEN MISINFORMATION IN MEMORY AFFECTS LATER INFERENCES. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(6), 1420-1436. doi:10.1037//0278-7393.20.6.1420
Johnson-Laird, P. N. (2012). Inference with mental models. *The Oxford handbook of thinking and reasoning*, 134-145.
Khazanov, L., & Prado, L. (2010). Correcting Students' Misconceptions about Probability in an Introductory College Statistics Course. *Adults Learning Mathematics, 5*(1), 23-35.
Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213-218. doi:10.1177/00131640121971185
Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences*: American Psychological Association.
Kowalski, P., & Taylor, A. K. (2009). The Effect of Refuting Misconceptions in the Introductory Psychology Class. *Teaching of Psychology, 36*(3), 153-159. doi:10.1080/00986280902959986
Krauss, S., & Wassner, C. (2002). *How significance tests should be presented to avoid the typical misinterpretations.* Paper presented at the Proceedings of the Sixth International Conference on Teaching Statistics. Cape Town, South Africa: International Association for Statistics Education. Online: www. stat. auckland. ac. nz/~ iase/publications.
Krueger, J. I., & Heck, P. R. (2019). Putting the P-Value in its Place. *American Statistician, 73*, 122-128. doi:10.1080/00031305.2018.1470033
Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155-177. doi:10.3758/s13423-017-1272-1
Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC research notes, 8*(1), 84.
Lecoutre, M. P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology, 38*(1), 37-45.
Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131. doi:10.1177/1529100612451018
Lipson, K. (2002). *The role of computer based technology in developing understanding of the concept of sampling distribution.* Paper presented at the Proceedings of the Sixth International Conference on Teaching Statistics. Cape Town South Africa.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232-257. doi:10.1016/j.cogpsych.2006.09.006

Macher, D., Papousek, I., Ruggeri, K., & Paechter, M. (2015). Statistics anxiety and performance: blessings in disguise. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.01116

McGarr, O. (2009). A review of podcasting in higher education: Its influence on the traditional lecture. *Australasian Journal of Educational Technology, 25*(3), 309-321.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*: Yale University Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: Methodological paradox. *Philosophy of Science, 34*(2), 103-115. doi:10.1086/288135

Murtonen, M. (2005). University students' research orientations: Do negative attitudes exist toward quantitative methods? *Scandinavian Journal of educational research, 49*(3), 263-280.

Murtonen, M., & Lehtinen, E. (2003). Difficulties experienced by education and sociology students in quantitative methods courses. *Studies in Higher Education, 28*(2), 171-185. doi:10.1080/0307507032000058064

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior, 32*(2), 303-330. doi:10.1007/s11109-010-9112-2

Pacer, M., & Lombrozo, T. (2017). Ockham's Razor Cuts to the Root: Simplicity in Causal Explanation. *Journal of Experimental Psychology-General, 146*(12), 1761-1780. doi:10.1037/xge0000318

Peachey, C. (2019). [Comparison of the efficacy of written and video corrections]. Unpublished raw data.

Pearson, K. (1900). *The grammar of science*. London: A & C Black. (Original work published 1892)

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.00223

Peter, C., & Koch, T. (2016). When Debunking Scientific Myths Fails (and When It Does Not): The Backfire Effect in the Context of Journalistic Coverage and Immediate Judgments as Prevention Strategy. *Science Communication, 38*(1), 3-25. doi:10.1177/1075547015613523

Qualtrics (2019). Qualtrics [Computer software]. Retrieved from https://www.qualtrics.com/uk/.

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition, 35*(8), 2019-2032. doi:10.3758/bf03192934

Rothman, K. J. (2014). Six Persistent Research Misconceptions. *Journal of General Internal Medicine, 29*(7), 1060-1064. doi:10.1007/s11606-013-2755-z

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi:10.3758/pbr.16.2.225

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in experimental social psychology, 39*, 127-161.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? *Psychology of Learning and Motivation: Advances in Research and Theory, 41*, 265-292. doi:10.1016/s0079-7421(02)80009-3

Simmons, J. P., Nelson, L. D., &

Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98-113.

Tenney, E. R., Cleary, H., & Spellman, B. A. (2009). The Other Dude Did It: A Test of the Alternative Explanation Defense. *Jury Expert, 21*, 37.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years. *Psychological Methods, 22*(2), 217-239. doi:10.1037/met0000100

Vittengl, J. R., Bosley, C. Y., Brescia, S. A., Eckardt, E. A., Neidig, J. M., Shelver, K. S., & Sapenoff, L. A. (2004). Why are some undergraduates more (and others less) interested in psychological research? *Teaching of Psychology, 31*(2), 91-97. doi:10.1207/s15328023top3102_3

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804. doi:10.3758/bf03194105

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181-207): Springer.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58-76. doi:10.3758/s13423-017-1323-7

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57. doi:10.3758/s13423-017-1343-3

Weisman, K., & Markman, E. M. (2017). Theory-based explanation as intervention. *Psychonomic Bulletin & Review, 24*(5), 1555-1562. doi:10.3758/s13423-016-1207-2

Williams, J. L., McCutcheon, L. E., Fava, D. A., & Aruguete, M. S. (2017). Using Research on Celebrity Admiration to Enhance your Statistics Class. *North American Journal of Psychology, 19*(3), 585-600.

Wilson, E. A. H., Makoul, G., Bojarski, E. A., Bailey, S. C., Waite, K. R., Rapp, D. N., . . . Wolf, M. S. (2012). Comparative analysis of print and multimedia health materials: A review of the literature. *Patient Education and Counseling, 89*(1), 7-14. doi:10.1016/j.pec.2012.06.007

Wilson, E. A. H., Park, D. C., Curtis, L. M., Cameron, K. A., Clayman, M. L., Makoul, G., . . . Wolf, M. S. (2010). Media and memory: The efficacy of video and print materials for promoting patient education about asthma. *Patient Education and Counseling, 80*(3), 393-398. doi:10.1016/j.pec.2010.07.011

Wood, T., & Porter, E. (2019). The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior, 41*(1), 135-163. doi:10.1007/s11109-018-9443-y

Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*: University of Michigan Press.