Journal of Business and Psychology (2024) 39:137–157
https://doi.org/10.1007/s10869-022-09862-8

ORIGINAL PAPER

Communicating Adverse Impact Analyses Clearly: A Bayesian Approach

Karyssa A. Courey[1] · Frederick L. Oswald[1]

Accepted: 15 November 2022 / Published online: 2 December 2022

Abstract
Adverse impact results from company hiring practices that negatively affect protected classes. It is typically determined on the basis of the 4/5ths Rule (which is violated when the minority selection rate is less than 4/5ths of the majority selection rate) or a chi-square test of statistical independence (which is violated when group membership is associated with hiring decisions). Typically, both analyses are conducted within the traditional frequentist paradigm, involving null hypothesis significance testing (NHST), but we propose that the less-often-used Bayesian paradigm more clearly communicates evidence supporting adverse impact findings, or the lack thereof. In this study, participants read vignettes with statistical evidence (frequentist or Bayesian) supporting the presence or absence of adverse impact at a hypothetical company; then they rated the vignettes on their interpretability (i.e., clarity) and retributive justice (i.e., deserved penalty). A Bayesian analysis of our study results finds moderate evidence in support of no mean difference in either interpretability or retributive justice, across three out of the four vignettes. The one exception was strong evidence supporting the frequentist vignette indicating no adverse impact being viewed as more interpretable than the equivalent Bayesian vignette. Broad implications for using Bayesian analyses to communicate adverse impact results are discussed.

Keywords Bayesian analysis · Statistical paradigms · Adverse impact · Statistical communication

Adverse Impact

One of the most groundbreaking legal decisions advancing equity in the workplace was Title VII of the Civil Rights Act of 1964, which prohibits employers from discriminating against employees or job applicants on the basis of sex, race, color, national origin, and religion (Civil Rights Act, 1964). More explicitly, Title VII protects against two types of discrimination: disparate treatment and disparate impact. *Disparate treatment* refers to intentional discrimination against an individual. A clear example of disparate treatment would be requiring only women, and not other applicants, to perform a physical task as part of a job application. In contrast, *disparate impact*, a term often used interchangeably with adverse impact, describes when an employment practice or policy has more of a negative effect against a protected class. Adverse impact is often measured using the 4/5ths Rule (and/or a statistical significance test) as outlined in the EEOC's *Uniform Guidelines on Employee Selection Procedures* (*UGESP*; 1978).

The 4/5ths Rule specifies that a selection rate for a particular group defined by race or sex, for example, should be at least 4/5ths (80%) that of the majority group (*UGESP*, 1978). A statistical term that is often used in this context is the *impact ratio*, which is the ratio of the subgroup selection rates. Although it is possible to apply a confidence interval to the impact ratio directly and thus understand whether it violates the 4/5ths Rule (Morris & Lobsenz, 2000), another common approach is to conduct a chi-square test of statistical independence, where the null hypothesis is that hiring rates are equal between subgroups (Bobko & Roth, 2004). In following these two standards, when the 4/5ths Rule is violated and the chi-square independence test is statistically significant, one can infer the presence of adverse impact.

Although a Bayesian approach is much less commonly used when determining the presence or absence of adverse impact, it carries a number of potential advantages for conducting analyses, while also providing results that are potentially easier to communicate statistical results more clearly.

✉ Karyssa A. Courey
kac23@rice.edu

[1] Department of Psychological Sciences, Rice University, 6100 Main Street - MS25, Houston 77005, USA

We first briefly review Bayesian methods to provide important introductory context for our study about communicating statistical results from adverse impact analyses.

Bayesian Methods

Virtually, all organizational research findings are based on traditional null hypothesis significance testing [NHST], which falls under the frequentist paradigm for statistical analysis. Bayesian statistics is another popular statistical paradigm with fundamentally different views about truth, probability, data, and parameters (Jebb & Woo, 2015; Kruschke et al., 2012; Zyphur & Oswald, 2015). Although some studies in organizational research have used Bayesian analyses, there are relatively few of them (e.g., Ballard et al., 2018; Grand, 2017; Jackson et al., 2016).

Some key terms are important to many Bayesian analysis: e.g., the *prior distribution*, the *likelihood*, the *posterior distribution*, and the *Bayes factor*. To convey the meaning of these terms, we will use the two-sample Bayesian t-test on standardized data as a running example (see Gronau et al., 2019). Consider Fig. 1, which portrays three distributions of mean differences between the two groups involved in a Bayesian t-test. The shaded blue region represents the *prior distribution*, which specifies the distribution of mean differences in a model that we expect before observing the data. Here, you can see that the prior is centered at zero and contains a wide range of mean differences. This prior distribution is represented as P(H), or the probability of the hypothesis, before observing any data.

The *likelihood* is the shaded red region on the right, and it tells us the range of probabilities of observing the mean differences that we did, based on the data and model of interest. Most people have heard of *maximum likelihood* estimation, which is actually the peak of the entire likelihood distribution, which in this case is about a 0.40 standard deviation unit separation of the group means. The entire likelihood distribution is represented as P(D | H), or the probability of the data, given the hypothesis.

The *posterior distribution* is the shaded green region and is proportional to the prior being updated (multiplied) by the likelihood distribution, where the data were observed. Because the updating is a multiplicative process, the posterior distribution is an average of the prior and the likelihood distributions, such that the posterior is always situated somewhere between the prior and the likelihood. Here, the

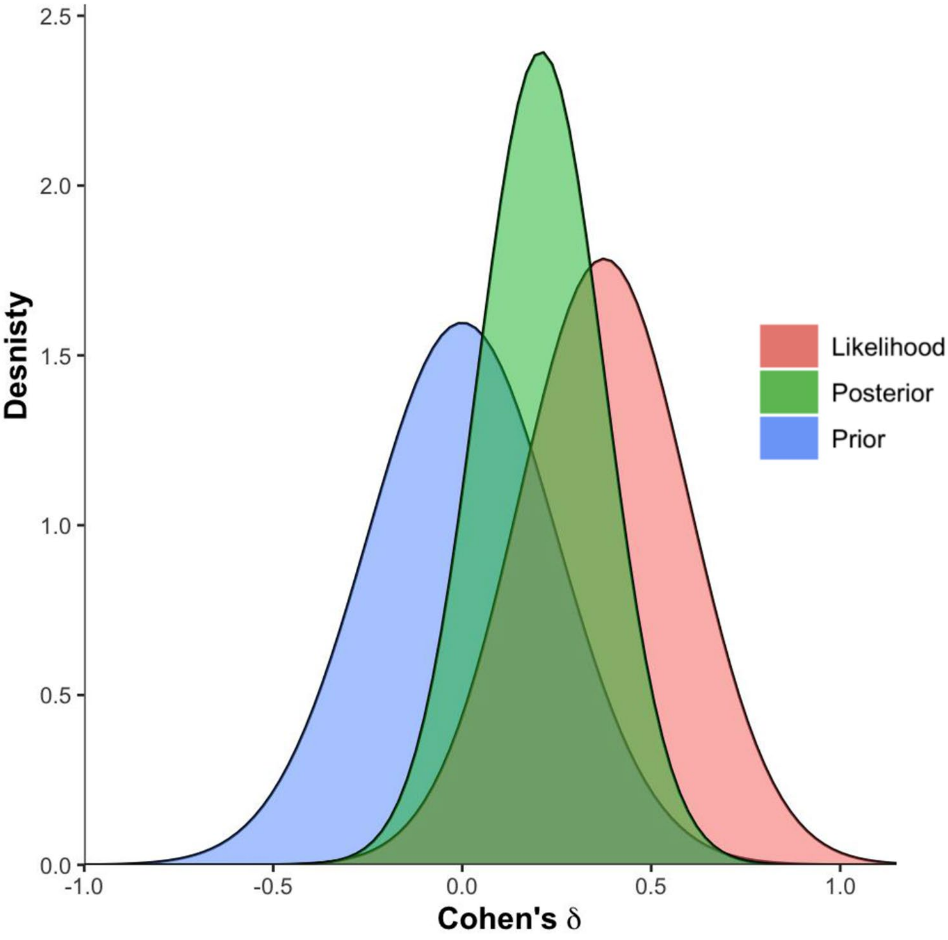


Fig. 1 Mean differences between two groups: prior (blue, on the left), likelihood (red, on the right), and posterior distribution (green, in the center). The posterior distribution lies between the prior and the likelihood. Given these are Cohen's δ effect sizes (the population version of Cohen's d), 0.2 to 0.5 would be considered a small effect, 0.5 to 0.8 a medium effect, and greater than 0.8 a large effect, by conventional standards (Cohen, 1988)

posterior distribution has a modal value of about 0.20, which is a mean difference located between the prior and posterior modal values; it constitutes a small effect size (Cohen's δ = 0.20; Cohen, 1988). Formally, the posterior distribution is the probability of a hypothesis given the data, and it is represented as P(H | D).

The posterior distribution reflects a "compromise" between the likelihood and the prior, such that the more data we have, the more the posterior distribution reflects the likelihood (and the prior has less influence), and conversely, the less data we have, the more the posterior reflects the prior distribution (and the likelihood has less influence). Bayes Theorem is what helps us find the most appropriate mathematical "compromise," where again, the prior distribution gets updated (multiplied) by the likelihood (data and model) to produce the posterior distribution. Mathematically,

$$P(H|D) = \frac{P(H) * P(D|H)}{P(D)} \tag{1}$$

Note that P(D) in the denominator is simply a normalizing constant (i.e., the numerator does the aforementioned updating, and P(D) simply makes the posterior a probability distribution so that the area under the curve is 1). Because of this, people often represent Bayes Theorem in Eq. 1 as:

$$P(H|D)\ \alpha\ P(H) * P(D|H) \tag{2}$$

where α means "is proportional to." That way, we see more clearly that the posterior distribution is proportional to the prior times the likelihood. From this posterior distribution, researchers then obtain and report the most probable parameters. Then, the range of this distribution is summarized using another widely used term in Bayesian analysis: the 95% *credible interval*.

Now we turn to the *Bayes factor*, which indicates how well the alternative hypothesis, given the data (numerator) explains the data relative to the null hypothesis, given the data (denominator). This is represented in Eq. 3, where we expand the numerator and denominator using Bayes Theorem, such that the ratio of posterior distributions is equal to the ratio of priors, multiplied by the Bayes factor (or the ratio of marginal likelihoods):

$$\begin{gathered}\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(H_1)}{P(H_0)} * \frac{P(D|H_1)}{P(D|H_0)} \\ \text{posterior odds} = \text{prior odds} * \text{BF}\end{gathered} \tag{3}$$

Notice that the prior is the ratio involving $P(H_1)$ and $P(H_0)$, such that if we were to assign the same prior distribution to each model, the ratio of priors (conditional across all possible mean differences) is simply 1, making the priors unnecessary and the Bayes factor (BF) the same as the ratio of the posterior probabilities (van Ravenzwaaij & Etz, 2021):

$$\begin{gathered}\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \\ \text{posterior odds} = \text{BF}\end{gathered} \tag{4}$$

This is not an unusual practice so that the data "speak for themselves," such that the Bayes factor tells us directly how much the evidence supports the alternative versus null models.

Table 1 shows common rule-of-thumb language that researchers use for interpreting the Bayes factor as an indicator of relative probabilities for evidence in support of the null vs. alternative hypothesis. To provide two numerical examples: (1) A BF_{10} of 5 (or BF_{01} of 1/5 or 0.20) would be described as moderate evidence in support of the alternative hypothesis (subscript = 1), relative to the null hypothesis (subscript = 0). (2) A BF_{10} of 1/15 or 0.07 (or BF_{01} of 15) would indicate that the null hypothesis provides strong evidence for explaining the data, relative to the alternative hypothesis. Importantly, although we used the subscripts to refer to the alternative model and the null model, they could be used to indicate and compare any two models more generally.

Bayesian Advantages for Communicating Adverse Impact

Having provided some context for conducting Bayesian analyses, we can now turn to the central thesis of the current work: How Bayesian analyses stand to improve upon frequentist analyses when identifying and communicating adverse impact statistics. For decades, scholars have lamented the many problems with using *p* values in practice (e.g., Goodman, 2008; Greenland et al., 2016; Hoekstra et al., 2006). In fact, the American Statistical Association (ASA) released an official statement on *p* values, advising against the sole use of *p* values for testing the truth of a

Table 1 Evidence provided by the Bayes factor

BF_{10} (H_0 over H_1)	Evidence[a]	BF_{10} (H_1 over H_0)
0.33—1	Anecdotal	1—3
.33-.1	Moderate	3—10
0.1—0.03	Strong	10—30
.03-.01	Very Strong	30—100
<.01	Extreme	>100

The leftmost column provides evidence in terms of the null hypothesis being better able to explain the data relative to the alternative hypothesis. The rightmost column provides equivalent evidence in terms of the alternative hypothesis being better able to explain the data relative to the null hypothesis.

[a] Adapted from Lee and Wagenmakers's (2013) revision of Jeffreys's (1961) classification.

hypothesis, for indicating effect size, or for reaching scientific conclusions, while at the same time recommending Bayesian methods as a potential way to focus on estimation rather than testing (Wasserstein & Lazar, 2016).

Using a Bayesian approach comes with a number of potential advantages: communicating uncertainty in adverse impact findings in a more understandable manner; informing the data with various sources of prior knowledge if desired (e.g., meta-analytic findings, plaintiff vs. defendant beliefs); directly supporting the null hypothesis of no adverse impact when such evidence is found; and having greater flexibility in model design. Each of these points is addressed in turn below.

Natural Interpretation

First, when communicating statistical findings on adverse impact, it is important to be clear without compromising technical accuracy, and to that end, Bayesian methods lend themselves to a more natural and intuitive interpretation of results. For example, a 95% credible interval on the standardized mean difference (effect size) from a Bayesian *t*-test can be described in a natural and straightforward statement such as "95% of the most probable population effect sizes fall between 0.53 and 0.74, and the distribution has a median value of 0.64." This type of statement would not be technically correct when using a 95% confidence interval in a traditional frequentist analysis. Instead, to be true to the statistics, a frequentist must say, "there is a 95% chance that the interval from 0.53 to 0.74 captures the population effect size underlying the data, and our best estimate for that parameter is 0.64." The 95% confidence interval only *indirectly* indicates which other mean differences are possible, whereas the 95% credible interval is much more direct.

Prior research also supports that Bayesian results may be more positively perceived and understood than those from the traditional frequentist paradigm. For example, Chandler et al. (2020) instructed participants to choose between an old and a new educational software program, based on equivalent data presented under a frequentist or Bayesian framework (holding cost constant). Participants were more likely to support the new software program (and be more confident in their choice) when its effectiveness in improving achievement was presented under a Bayesian framework (i.e., a posterior probability distribution) versus under NHST (i.e., a 95% confidence interval). In a similar vein, Hurwitz (2020) presented legal aids with vignettes under the frequentist or Bayesian paradigm for an educational program that varied in cost and ease of implementation. This study also found that results presented in a Bayesian framework were more readily understood, and study participants were more likely to endorse the program. Similarly, in the present study, we are interested in understanding whether adverse impact statistical evidence is more interpretable when presented in a Bayesian versus a frequentist framework.

Incorporating Prior Information

The breadth and potential for using prior distributions for organizational analyses in I-O psychology have largely not been explored (e.g., Oswald et al., 2021). *Noninformative priors* are common and the typical default in software for many standard analyses (e.g., the prior we mentioned previously that assumes each hypothesis is equally likely). However, there are analyses where selecting an *informative prior* is sensible, based on past data and analyses (e.g., meta-analytic results on a similar research topic, previous research findings at a similar organization). Informative priors might also be introduced on a logical basis (e.g., in multilevel models, variance estimates cannot be negative, and unusually large values can be given smaller prior probabilities).

The use of informative priors is often viewed as an advantage of Bayesian analyses, compared with the traditional frequentist approach. Researchers often hold prior beliefs when they conduct a study, and this is not exclusive to the Bayesian framework; Bayesians just account for this belief explicitly (Kruschke, 2010). Before conducting a study, researchers engage in the research literature to locate the empirical evidence that supports their hypotheses, research design, and analysis of results; the difference is that a Bayesian analysis can formally incorporate this prior knowledge into the prior distribution (Greenland, 2006), and a frequentist typically cannot. For example, using meta-analytic data to serve as a prior distribution for a specific analysis has the potential to increase a local study's validity (Newman et al., 2007, p. 1406).

For the present study, we elected to use a noninformative prior under the assumption that we lack prior information about people's perceptions about the presence or absence of adverse impact; however, in future research, an informative prior distribution could incorporate past information from perceptions from relevant audiences who are judging adverse impact information in a particular case (e.g., judgments by organizational stakeholders, employees in the protected class, lawyers, and other subject matter experts). A future study could even use the results from the current study to contribute to an informed prior distribution. Note that the choice of a Bayesian prior is flexible and can reflect the strength of our beliefs, such that vaguer assumptions will lead to wider priors with larger variances, and stronger assumptions lead to narrow priors with smaller variances (e.g., van de Schoot et al., 2021). No matter the prior that is selected, Bayesian robustness checks are common, showing how sensitive the Bayes factor is to changes in the width of the prior distribution (Greenland, 2006; van Doorn et al., 2021). Less sensitivity indicates robustness, meaning that the choice of prior has less influence on the final results obtained from the posterior distribution.

Accumulating Evidence to Support the Null Hypothesis

Third, Bayesian analyses provide the ability to accumulate evidence to directly support the null hypothesis of no difference in the perception of adverse impact under frequentist or Bayesian paradigms (which can also be stated as accumulating evidence against perceptions of the presence of adverse impact). In contrast, in the frequentist framework, we can merely "fail to reject the null" of no difference between adverse impact perceptions, rather than support it (Dienes, 2014). In the Bayesian framework, using the Bayes factor and/or the ratio of posterior distributions, we can accumulate direct evidence in support of the alternative hypothesis, the null hypothesis, or conclude there is not enough evidence to support *either* hypothesis. Moreover, the Bayes factor allows us to quantify the amount of evidence that we have in support of the alternative versus null hypothesis (e.g., anecdotal, moderate, strong, extreme; see Table 1), rather than using the dichotomous thinking of *p* values. Bayesian analysis is directly relevant and useful for adverse impact cases, because we can test whether the evidence supports the plaintiff's or the defendant's perceptions of adverse impact in light of the data and model.

Represent Uncertainty in Estimation

Bayesian methods, through their definition of probability, represent uncertainty in the posterior distribution, with a wider distribution representing a greater range of probable parameter values, such as for the impact ratio in an adverse impact analysis. Thus, rather than definitively concluding that adverse impact is or is not present in an organization (as with a test of statistical significance) or providing the accuracy of a single estimate of adverse impact (as with a 95% confidence interval), we can instead state how much evidence supports the presence or absence of adverse impact (through the Bayes factor) and which population values are most probable (via the 95% credible interval), as the previous example demonstrated.

To more closely examine these statistical and communication benefits of Bayesian analyses, we conducted an experimental study containing vignettes that present statistical evidence (a) either supporting or not supporting adverse impact (a within-subjects manipulation) in (b) either the Bayesian or traditional frequentist framework (a between-subjects manipulation). For each vignette, we then measured participants' perceptions of adverse impact in terms of *interpretability*, reflecting their understanding of statistical evidence of adverse impact. We also measured their level of *retributive justice*, or their beliefs about the degree to which companies deserve penalties for having policies that contribute to adverse impact. Hypotheses that stem from this study design are specified below. Note that the statistical results of this study are presented in a Bayesian framework.

Measuring the Effective Statistical Communication of Adverse Impact

Interpretability

Given the widespread errors in statistical communication found in society, we were interested in examining the *interpretability* of both frequentist and Bayesian paradigms when they are presented in a manner that is typical and technically accurate. As mentioned above, a couple of recent studies (e.g., Chandler et al., 2020; Hurwitz, 2020) have observed that results presented in a Bayesian framework tend to be easier to understand, compared with similar results presented in a frequentist paradigm. Although these two studies explained parameter estimation when measuring interpretability, we were interested in examining if these findings extend to the interpretability of hypothesis testing for adverse impact cases.

> *Hypothesis 1*: Participants will tend to find adverse impact results presented in a Bayesian framework as being more interpretable than those presented in a frequentist framework, no matter the level of evidence.

Retributive Justice

In our experimental context, we also considered individual differences in retributive justice on judgments of adverse impact. *Retributive justice* is aptly defined by Wenzel and Okimoto (2016) as "the subjectively appropriate punishment of individuals or groups who have violated rules, laws, or norms and, thus, are perceived to have committed a wrongdoing, offense, or transgression" (p. 238). In other words, it involves proportional punishment for a committed crime while maintaining the rights of the innocent (Walen, 2021). Retributive justice is an important component of human decision-making as well as our legal system in the USA, although punishment for crimes often (if not always) incorporates contextual factors, rehabilitative goals, and other mitigating factors that weigh against strict proportionality in terms of actual punishment.

Turning to adverse impact, organizations that violate the 4/5ths Rule (perhaps accompanied by a statistically significant chi-square test of independence) might be expected to receive at least some form of penalty in a legal case, and we hypothesize that adverse impact results expressed in the Bayesian paradigm will be more compelling as degree-of-belief statements, thus leading to better calibrated penalties for companies that align with retributive justice notions.

> *Hypothesis 2*: Within vignettes that clearly reflect adverse impact, participants provided with Bayesian results will tend to provide higher ratings of retributive justice than participants provided with comparable frequentist results.
> *Hypothesis 3*: Within vignettes that clearly reflect a lack of adverse impact, participants provided with Bayesian results will tend to provide lower ratings of retributive justice than participants provided with comparable frequentist results.

Method

Participants

We recruited 120 participants via Prolific, which has been suggested to produce higher quality data relative to other crowdsourcing platforms, such as MTurk (Eyal et al., 2021; Peer et al., 2017), with the goal of capturing a sample that is representative of a typical jury. Specifically, participants were required to be at least 18 years old or older and report English as their first language. Participants were, on average, 39.9 years old (*SD* = 13.2; min = 18, max = 84). Table 2 contains additional demographic information about the study sample.

Stimuli

Our experimental vignettes provided (a) frequentist or Bayesian results, where (b) results either supported adverse impact or a lack thereof. Specifically, the following four vignettes were created (see Appendix 1): (1) frequentist chi-square test with $p = 0.002$ and a very large effect size, (2) Bayesian chi-square test with $BF_{10} = 38$ and very strong evidence favoring the alternative hypothesis, (3) frequentist chi-square test with $p = 0.995$ and a very small effect size, (4) Bayesian chi-square test with $BF_{01} = 30$ and very strong evidence favoring the null hypothesis. For example, Vignette 1 stated the following:

"Imagine that you are selected to serve on a jury during the trial of Smith v. Bright Light. Bright Light Company, which manufactures and sells lightbulbs, is on trial for adverse impact after a recent job candidate, Ms. Smith, claimed that she was not hired because of her gender. Ms. Smith's lawyer hired a subject matter expert, Dr. Williams, to present statistical evidence to you and the other jurors at the trial.

The prosecution calls Dr. Williams to the stand. In response to the prosecution's question about statistical evidence, Dr. Williams reports (a) the raw data, (b) results from applying the 4/5ths Rule, and (c) results from a chi-square test of independence as follows:

> "A total of 60 people applied (40 men, 20 women) and 25 people were hired (23 men, 2 women)."
> "In addition, the 4/5ths Rule indicates evidence of adverse impact, because 10% of the women and 57.5% of the men were hired, resulting in a ratio of .17, which is less than .80 (or 4/5ths)."
> "A chi-square test was used to determine if the observed hiring rate of men and women differed from a fair hiring rate (i.e., where the hiring rate is completely independent of gender). We observed the hiring rate of men compared to women being statistically significant (p = .002), with men being hired at a greater rate than women; thus, we can reject the null hypothesis that the hiring rate is independent from gender. This corresponds to a phi-coefficient of .39, which by conventional rules of thumb constitutes a very large effect size."

Within each pair of Vignettes 1 and 2 (evidence supporting adverse impact) and Vignettes 3 and 4 (evidence against adverse impact), each shared the same verbal description, raw numbers, and impact ratio; they only differed in their reporting of the chi-square test of independence (i.e., frequentist or Bayesian). Each participant viewed either two frequentist vignettes (1 and 3) or two Bayesian vignettes

Table 2 Participant descriptive statistics: gender, ethnicity, education, and statistics courses

Gender	Ethnicity	Education	Statistics courses
39% Male	80% White/Caucasian	18% High school degree or equivalent	43% 0 courses
56% Female	13% Multiethnic	22% Some college	32% 1 course
3% Non-binary	4% Asian/Asian American	8% Associate degree	17% 2 courses
1% Genderqueer	2% Other	35% Bachelor's degree	4% 3 courses
	1% Black/African American	15% Master's degree	1% 4 courses
		2% Doctorate	3% 5 or more courses

Total N = 104. Percentages are rounded to the nearest whole number. One participant did not specify gender. Multiracial ethnicity refers to participants who selected two or more options to identify their ethnicity.

(2 and 4) in this 2 (frequentist or Bayesian, between-subjects) × 2 (favoring adverse impact or no adverse impact, within-subjects) mixed design. Vignettes were presented in counterbalanced order. Dependent variables were responses to the interpretability and retributive justice scales, described below.

Procedure

The entire experiment was hosted on Qualtrics, a web-based survey platform. After consenting to participate in the study, the participants were presented with information on what adverse impact is and how it is typically measured, as one might when instructing a jury about the nature of a case. Next, participants answered one content check question on the information they just read and were provided feedback. Then, participants were instructed to imagine that they were serving on a jury and were being asked to evaluate the statistical evidence being presented by a subject matter expert. After reading the vignette, the participants were asked a series of questions regarding the interpretability of the evidence and to what extent the company deserved retributive justice. Finally, participants responded to the demographic questionnaire, indicated the number of statistics courses they have completed, answered two questions regarding their general statistics and Bayesian statistics knowledge, and finally, completed a subjective numeracy scale. The entire experimental task lasted about 10 min.

Measures

Interpretability of Evidence

We used a 4-item composite to assess the ability of the participant to understand the statistical evidence presented in the vignette they read. These items are as follows: "The evidence was easy to understand," "The evidence made sense," "I could interpret the results easily," and "The evidence was clearly communicated." Responses were on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). The measure had high internal consistency reliability within both vignettes ($\alpha = 0.95$), with high average inter-item correlations of 0.83–0.85. Reliability analyses were conducted using the R package psych (Revelle, 2016).

Retributive Justice Scale

We used van Prooijen and Coffeng's (2013) 5-item measure of retributive justice. This measure was necessarily adapted to the vignettes by (a) using the words *company* and *penalty* (instead of *offender* and *punishment*, respectively); (b) phrasing the questions as statements; and (c) changing the response scale to 1 = strongly disagree, 7 = strongly agree (instead of 1 = very mild punishment, 7 = very severe). The five items presented were as follows: ''The company should be penalized," ''The company deserves to be penalized," ''A penalty would be considered fair," ''A penalty would be considered justified," and ''A penalty would be considered appropriate." The measure has high internal consistency reliability within both vignettes ($\alpha = 0.96$–0.98) and high average inter-item correlations of 0.88–0.93.

Self-Rated Statistics Knowledge

To understand the sample in terms of their self-rated statistics knowledge, we asked three questions: "How many statistics courses have you taken?" (1, 2, 3, 4, 5, or more), "Please rate your knowledge of statistics, generally speaking" (0 = None, 5 = Average, 10 = Excellent), and "Please rate your knowledge of Bayesian methods, in particular" (0 = None, 5 = Average, 10 = Excellent). The average rating of statistics generally was 3.8 ($SD = 2.3$), and the average rating of Bayesian methods in particular was 1.3 ($SD = 2.1$), noting that both average values are below the scale midpoint of 5, meaning that statistics knowledge tended to be low in our sample.

Subjective Numeracy

We used Fagerlin et al.'s (2007) 8-item Subjective Numeracy Scale (SNS). A sample item is "When reading the newspaper, how helpful do you find tables and graphs that are parts of a story?" (1 = not at all helpful, 6 = extremely helpful). Participants' SNS scores are their average across the 8 items, with item 7 reverse coded. The average SNS composite score was 4.7 ($SD = 2.6$), above the scale midpoint of 3.5.

Results

Outlier Analysis

Earlier pilot testing suggested that we exclude any participants who spent less than 4 min completing the survey, because it would be nearly impossible to have read all the material carefully in that short period of time. Fifteen participants spent less than 4 min on the study and were thus removed from the analysis. An additional participant was removed for not consenting to the study. Four values were considered univariate outliers because their magnitude was greater than three standard deviations from the mean; we converted these values to the value at the third standard deviation. In the end, 104 total participants were included in the data analysis, with 54 assigned to the frequentist condition, and 50 assigned to the Bayesian condition.

Hypothesis Testing

To test our hypotheses, we first computed a Bayesian *t*-test comparing the frequentist and Bayesian paradigms when the corresponding vignettes had evidence supporting adverse impact. We then conducted a similar test for the corresponding vignettes with evidence against adverse impact. These two Bayesian *t*-tests were conducted for both the interpretability measure and for the retributive justice measure (i.e., four *t*-tests total). For all Bayesian *t*-tests, we used JASP statistical software (JASP, 2022), which applies a Cauchy default prior with a mean of zero and scale parameter of 0.707 for the effect size of the alternative hypothesis (note that this is the same as the default prior in the BayesFactor package in R; Morey & Rouder, 2021). This Cauchy distribution is symmetric and unimodal, similar to a univariate normal distribution, but with heavier tails (see Fig. 6 Appendix 2). By using this Cauchy distribution, we are specifying that the null hypothesis predicts an effect size of zero, and the alternative hypothesis reflects an effect size distribution centered at zero with an interquartile range of − 0.707 to 0.707 (see Wagenmakers et al., 2018). More simply, this is a noninformative prior that essentially places equal weight on the null and alternative hypotheses, such that neither hypothesis is favored before observing the data (and calculating the likelihood).

Table 3 Descriptive statistics: Interpretability and retributive justice across paradigms

		Interpretability		Retributive justice	
		Support for adverse impact	Against	Support for adverse impact	Against
Frequentist	Mean	5.14	5.32	5.55	1.71
	SD	1.41	1.30	1.52	1.00
Bayesian	Mean	5.26	4.38	5.24	1.81
	SD	1.27	1.72	1.37	1.06

Total $N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). *SD* = standard deviation.

Table 3 contains descriptive statistics and Table 4 summarizes the *t*-test results. Regarding the interpretability of vignettes that provided evidence supporting adverse impact, we observed moderate support for no difference between the Bayesian and frequentist paradigms ($BF_{01} = 4.37$, Cohen's $\delta = 0.08$, 95% CrI [− 0.45, 0.28]; see Fig. 6 Appendix 2 for all prior/posterior plots). The 95% CrI can be interpreted as 95% of the most probable values of the effect size fall between − 0.45 and 0.28, and the distribution has a median value of 0.08 (a small effect by the convention of Cohen, 1988). However, for the vignettes with evidence against adverse impact, we observed strong support for the frequentist paradigm being viewed as more interpretable ($BF_{10} = 16.81$, $\delta = 0.57$, 95% CrI [0.19, 0.97]; see Fig. 2), with a medium effect size by Cohen's (1988) conventions.

Regarding retributive justice, we observed moderate evidence supporting that the null hypothesis of no difference was better at explaining the data across paradigms, regardless of whether vignettes provided evidence supporting adverse impact ($BF_{01} = 4.81$, $\delta = 0.02$, 95% CrI [− 0.35, 0.38]) or against adverse impact ($BF_{01} = 4.32$, $\delta = 0.09$, 95% CrI [− 0.45, 0.28]; see Fig. 3). In other words, the paradigm for communicating results did not affect ratings of retributive justice.

To examine how much the observed Bayes factors change across priors that vary in their width (uncertainty), we computed a Bayesian Robustness Check (Fig. 7 Appendix 3). Note that the alternative model prior is being varied, whereas the null model prior remains the same. Results are said to be robust if the Bayes factor remains consistent across reasonable priors that vary in their width. Indeed, this is what we find in support of the null hypothesis in three cases: The null is supported for interpretability ratings in vignettes supporting

Table 4 *T*-test results: Interpretability and retributive justice scores across paradigms

	Interpretability		Retributive justice	
	Support for adverse impact	Against	Support for adverse impact	Against
BF_{01} (Null vs. Others)	4.37	0.06	4.81	4.32
$BF_{10} = 1 / BF_{01}$ (Others vs. Null)	0.23	16.81	0.21	0.23
Cohen's δ	0.08	0.57	0.02	0.09
95% CrI	[− 0.45, 0.28]	[0.19, 0.97]	[− 0.35, 0.38]	[− 0.45, 0.28]

Total $N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). Cohen's δ = Cohen's delta, the population standardized difference. 95% CrI = 95% credible interval. As a rule of thumb, $BF_{01} > 3$ provides moderate support for the null hypothesis of no mean difference between groups. In the second row, results are provided in terms of the alternative hypothesis such that $BF_{10} > 10$ provides strong support for the alternative hypothesis of a mean difference between paradigms. Note that row two is redundant because it is the exact reciprocal of row one (e.g., 4.37 = 1/0.23), but it may help with interpretability. All model error percentages (i.e., the error of the Gaussian quadrature) were very low, much less than the 20% rule of thumb (see van Doorn et al., 2021).

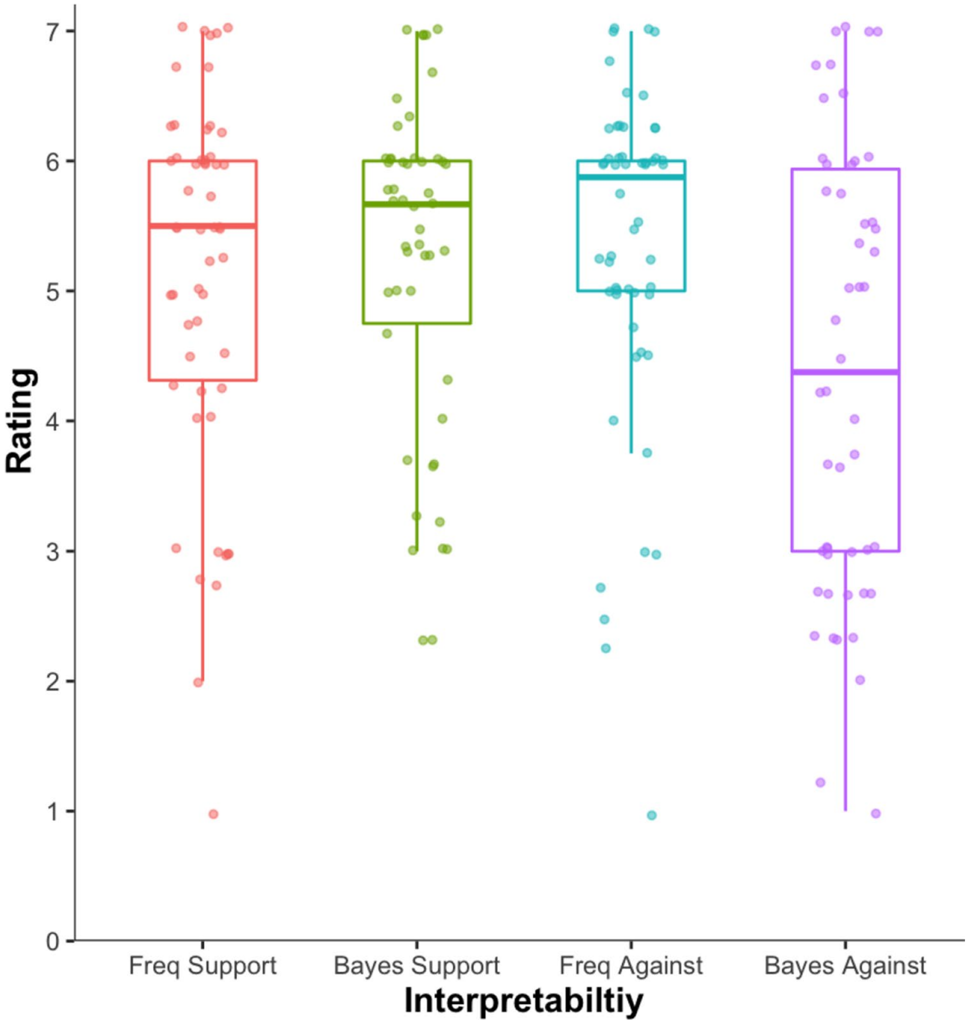


Fig. 2 Interpretability ratings across adverse impact vignettes. Total $N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). Interpretability scores are similar across paradigms, except for Bayesian paradigms with evidence against adverse impact, in which there is more variability in ratings. Data are jittered so that all data points are visible

adverse impact, and the null is also supported for retributive justice responses across paradigms, both when vignettes are supporting and against adverse impact. The alternative hypothesis is only supported for interpretability ratings following vignettes with evidence against adverse impact, such that the frequentist paradigm is viewed as more interpretable in this case (although both means are relatively high).

Bayesian Mixed-Model ANOVA

Second, as a more integrative statistical analysis that follows up on the Bayesian *t*-tests, we conducted two mixed-model Bayesian ANOVAs to explore between and within-subject effects for each measure. The first integrative analysis examined a 2 (paradigm: frequentist or Bayesian; between-subjects) × 2 (measure: interpretability support/against; within-subjects) Bayesian mixed-model ANOVA. We were interested in understanding how interpretability ratings differed across paradigms, within-participants. We applied the default prior in JASP of equal probabilities for each model (Rouder et al., 2012). Default priors should be easy to use, generalizable, and fit most cases in experimental psychology; the ANOVA default prior is an extension of the Cauchy prior (see Rouder et al., 2012). Table 5 shows that the "Measure + Paradigm + Measure*Paradigm" model (i.e., the interaction model) is 11.42 times better at explaining the data relative to the null model (i.e., the grand mean only; frequentist partial $\eta^2 = 0.09$). Note that we provide a frequentist partial η^2 because JASP does not currently provide effect sizes for Bayesian ANOVAs. This is a medium effect (frequentist partial $\eta^2 = 0.06$) based on Cohen's (1988) conventions.

The interaction model is also over 10 times better at explaining the data relative to all other models that were compared (see Table 7 Appendix 4 for model comparison between the best model, indicated by the $BF_{01} = 1$, and all other models). All error percentages were much less than the 20% rule of thumb (see van Doorn et al., 2021) which indicates that the observed Bayes factors are stable. Figure 4 shows the interaction between responses on the interpretability measure and the paradigm. Specifically, responses on the interpretability measure were similar following vignettes with evidence supporting adverse impact across paradigms.

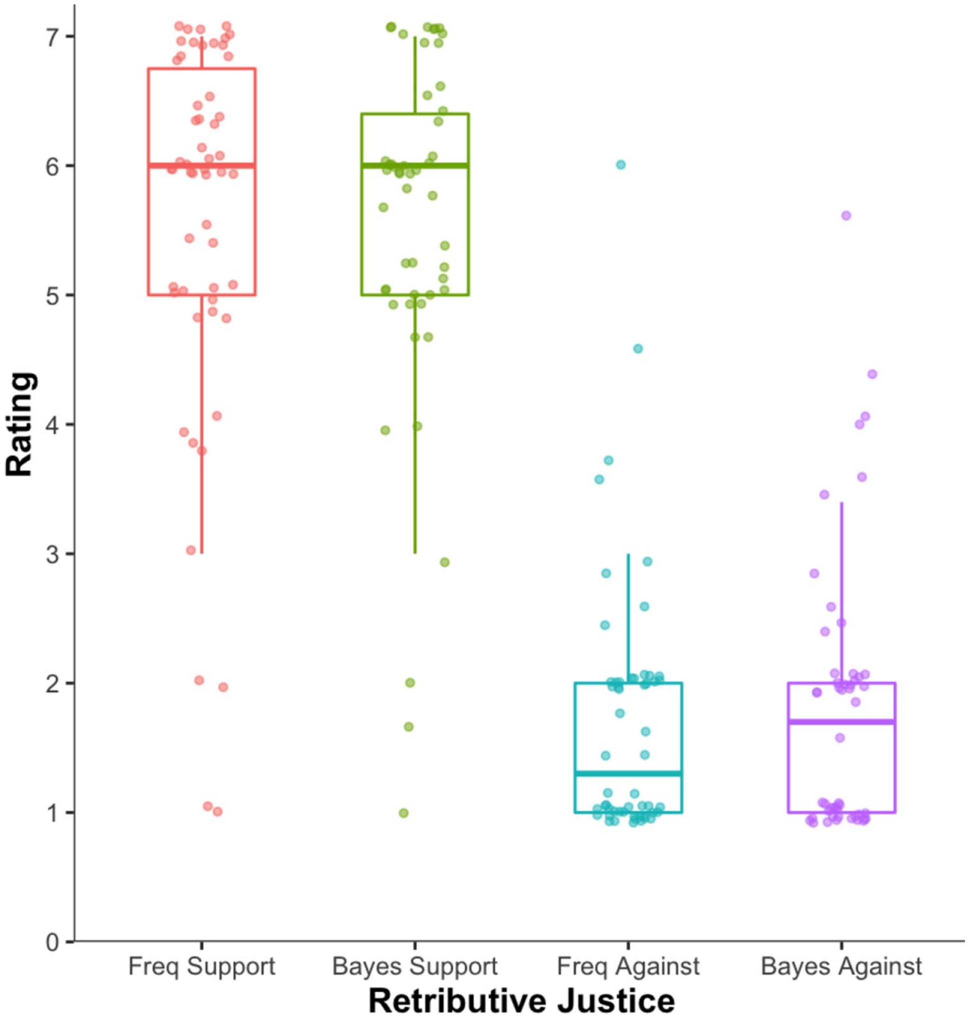


Fig. 3 Retributive justice ratings across adverse impact vignettes. Total $N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). Ratings on the retributive justice scale are similar across both paradigms, with much higher ratings following vignettes with evidence supporting adverse impact. Data are jittered so that all data points are visible

Table 5 Interpretability measure: Model comparison relative to the null model

Models	P(M)	P(M|data)	BF_M	BF_{10} (Others vs. Null)
Null model (Grand mean)	0.20	0.07	0.29	1.00[a]
Best model (Measure + Paradigm + Measure*Paradigm)	0.20	0.78	13.75	11.42[b]
Paradigm	0.20	0.06	0.24	0.84[c]
Measure	0.20	0.06	0.23	0.82[c]
Measure + Paradigm	0.20	0.05	0.19	0.66[c]

Total N = 104 (*n* = 54 frequentist; *n* = 50 Bayesian). *Measure* = interpretability measure; *P(M)* = prior probability; *P(M|data)* = posterior model probability; BF_M = posterior model odds; BF_{10} = Bayes factor of each model compared to the null model (other = 1, null = 0); numbers greater than 1 indicate support for other models and numbers less than 1 indicate support for the null model. All error percentages were less than 20% (see van Doorn, 2021).

[a] All possible models compared to the null model of no mean difference (row 1; column BF_{10}); the null model being compared to itself yield a BF_{10} of 1.

[b] The Measure + Paradigm + Measure*Paradigm model is considered the best model because it has the largest Bayes factor relative to the null model.

[c] The Paradigm model, Measure model, and Measure + Paradigm model are 0.84, 0.82, and 0.66 times as good at explaining the data relative to the null model, respectively.

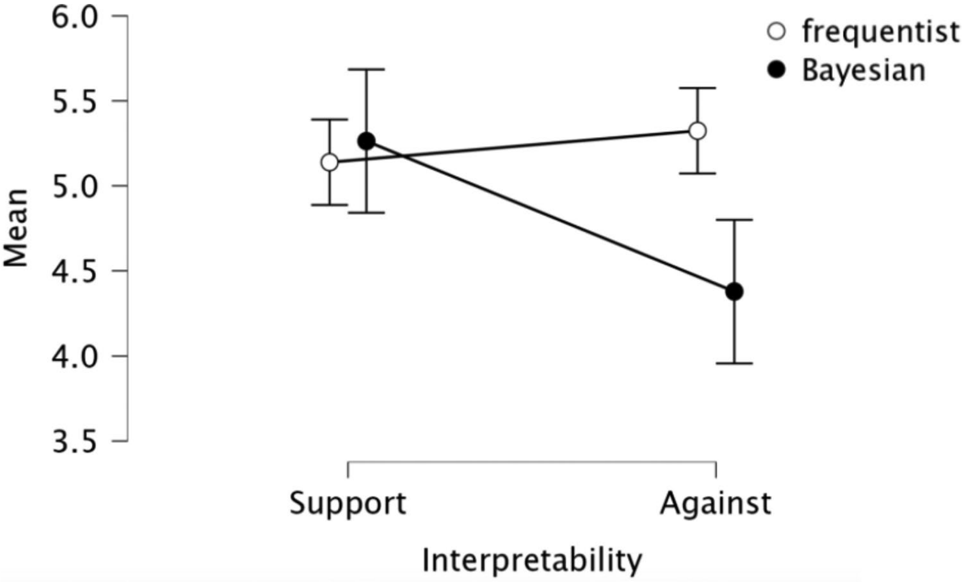


Fig. 4 Posterior distributions for interpretability of adverse impact vignettes. Total $N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). Descriptive plots for both interpretability measures (interpretability after paradigms supporting or against adverse impact), separated by paradigm. Vertical bars represent the 95% credible interval. The lines overlap, indicating that interpretability scores are relatively consistent between paradigms, supporting adverse impact; however, interpretability scores were much higher after the frequentist against adverse impact vignette compared to the Bayesian against adverse impact vignette

However, responses on the interpretability measure were higher following frequentist vignettes with evidence against adverse impact, compared to their Bayesian counterparts.

The second integrative analysis examined a 2 (paradigm: frequentist or Bayesian; between-subjects) × 2 (measure: retributive justice support/against; within-subjects) Bayesian mixed-model ANOVA. We applied the default prior in JASP of equal probabilities for each model.

Table 6 shows that relative to the null hypothesis of no difference (column BF_{10}), all models (with the exception of the between-subjects model) produce support that is extreme (where "extreme" is the descriptor recommended for the level of the Bayes factor; see Table 1).

The best model is the "Measure" model because it has the largest Bayes factor relative to the null model (see Table 7 Appendix 4 for model comparison between the best model, indicated by the $\mathrm{BF}_{01} = 1$, and all other models). The "Measure" model is 5.72 times better at explaining the data relative to the "Measure + Paradigm" model. The "Measure" model is also 26.60 times better at explaining the data relative to

Table 6 Retributive justice: Model comparison relative to the null model

Models	P(M)	P(M|data)	BF_M	BF_{10} (Others vs. Null)
Null model (Grand mean)	0.20	< 0.001	< 0.001	1.00[a]
Best model (Measure)	0.20	0.83	18.82	> 100[b]
Measure + Paradigm	0.20	0.14	0.67	> 100[c]
Measure + Paradigm + Measure*Paradigm	0.20	0.03	0.13	> 100[c]
Paradigm	0.20	< 0.001	< 0.001	0.16[d]

$N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). *Measure* = retributive justice measure; *P(M)* = prior; *P(M|data)* = posterior model probability; BF_M = posterior model odds; BF_{10} = Bayes factor of each model compared to the null model (other = 1, null = 0); numbers greater than 1 indicate support for other models and numbers less than 1 indicate support for the null model. All error percentages were less than 20% (see van Doorn, 2021).

[a] All possible models compared to the null model of no mean difference (row 1; column BF_{10}); the null model being compared to itself yield a BF_{10} of 1.

[b] The Measure model is considered the best model because it has the largest Bayes factor relative to the null model. It is also about 6 times better than the Measure + Paradigm model and 26 times better than the Measure + Paradigm + Measure*Paradigm model at explaining the data (see Table 7 Appendix 4 for model comparison between the best model and other models).

[c] The Measure + Paradigm model and Measure + Paradigm + Measure*Paradigm model are over 100 times better at explaining the data than the null.

[d] The Paradigm model is 0.16 times as good at explaining the data relative to the null model.

the interaction model and over 100 times better at explaining the data (frequentist partial $\eta^2 = 0.81$; a very large effect; Cohen, 1988) relative to the "Null model" and "Paradigm" model. Thus, the model that considers the measure alone (excluding the paradigm or any interactions between paradigm and measure) provides strong evidence for explaining the data. In other words, knowing the paradigm does not contribute to predicting retributive justice ratings. Figure 5 shows that the retributive justice responses are consistent across paradigms supporting or against adverse impact.

Discussion

Clear and accurate communication of employment rates and adverse impact findings is critical to improving organizational effectiveness, job applicant fairness, and Title VII compliance. We hypothesized that Bayesian statistics, more than their frequentist counterparts, would provide clearer communication of statistical findings with regard to adverse impact or the lack thereof. We failed to support Hypothesis 1, as there were no mean differences in interpretability between Bayesian and frequentist paradigms vignettes supporting the presence of adverse impact. However, we did observe increased interpretability of the frequentist paradigm when evidence was presented against adverse impact compared to the Bayesian counterpart.

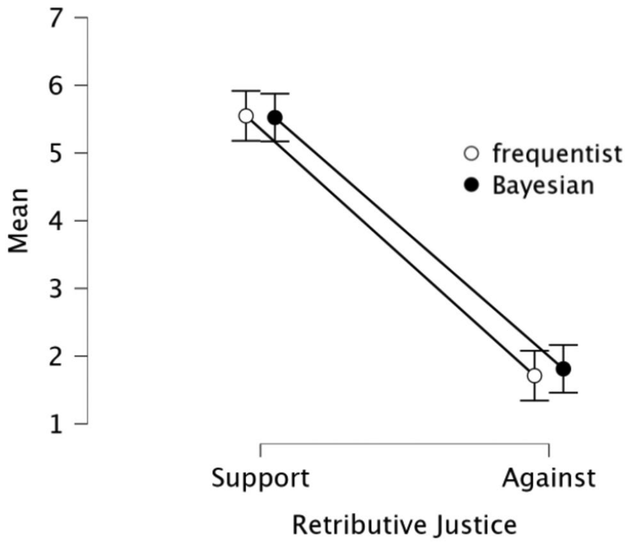


Fig. 5 Posterior distributions for retributive justice of adverse impact vignettes. Total $N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). Descriptive plots for both retributive justice measures (retributive justice after paradigms supporting or against adverse impact), separated by paradigm. The vertical bars represent the 95% credible interval. The lines are parallel and steadily decrease, indicating that retributive justice ratings are much higher following a vignette that provides evidence in support of adverse impact, no matter the paradigm

Furthermore, we examined whether these findings would be qualified by individual differences in the need for retributive justice (providing punishment in proportion to the employment offense). Contrary to Hypotheses 2 and 3, we failed to show that vignettes presenting evidence in a Bayesian framework were more likely to result in participants responding more (or less) severely on the retributive justice scale. Instead, our results support no mean difference on the retributive justice scale between the two paradigms. In this sense, the Bayesian paradigm was as effective as the frequentist paradigm for conveying statistical information about adverse impact.

For the most part, our results were comparable across paradigms, supporting no difference between frequentist or Bayesian ratings of retributive justice and interpretability (with the one exception of a paradigm difference in interpretability ratings for vignettes presenting evidence against adverse impact). This result should be viewed as both useful and promising, useful because Bayesian methods are at least no worse than frequentist methods when communicating results and making decisions about adverse impact, and promising because they offer additional advantages, such as accumulating and communicating evidence to support the null hypothesis. With Bayesian analysis, we can actually support no difference between groups (unlike NHST) and even quantify the amount of evidence with Bayes factors. By extension, we can also compare other probable alternative models, allowing for greater flexibility in model comparison and design.

Moreover, as we noted in the introduction, the Bayesian paradigm is worth turning toward for many advantages beyond the scope of the present study. For example, known information (e.g., previous company data on adverse impact) can be used to inform the model. With this, we can also examine the sensitivity of posterior distributions to the prior, which allows us to explore and understand how a range of prior distributions can affect the posterior. Such application of informative priors is a distinct advantage when handling small samples. Whereas most traditional adverse impact analyses on small samples result in a lack of statistical power, increasing Type II error (Morris, 2001), carefully selected Bayesian informed priors have the potential to improve the usability of small sample data and the quality of statistical outcomes (McNeish, 2016). To reiterate, informed priors have an even larger influence on the posterior when sample sizes are small, and thus one must select priors with care, comparing a range of options.

We attempted to extend another Bayesian advantage from Chandler et al. (2020) and Hurwitz (2020), that Bayesian methods are more readily understood, by investigating hypothesis testing across paradigms. It is possible that the hypothesis testing we presented may be more generally confusing for participants to grasp, compared to the probability distributions of parameters used by Chandler et al. (2020) and Hurwitz (2020). In fact, hypothesis testing and associated statistics, such as the chi-square test, are routinely misinterpreted among researchers of both the frequentist (e.g., Greenland et al., 2016) and Bayesian paradigms (e.g., Wong et al., 2021). In our study sample, interpretability ratings averaged around 5 (slightly agree) out of 7 (strongly agree) with great variability. This suggests that participants, on average, found statistical results only slightly interpretable overall. This is not entirely surprising because most of our study sample either had taken no statistics courses or they had taken one or two statistical courses that most likely were taught in the frequentist paradigm (Aiken, 1990; Kline, 2020), putting Bayesian statistics at an inherent disadvantage in our study.

The difficulty in understanding hypothesis testing, especially for people who are not knowledgeable of statistics (i.e., our sample also self-rated below average in terms of statistics knowledge), may lend an understanding of why the frequentist vignette presenting evidence against adverse impact was more interpretable than the Bayesian counterpart. It is possible that people skimming the vignette found the statement "a very weak effect" more interpretable than "strong evidence supporting that hiring rates do not differ by gender." This suggests that we, perhaps, should have revised our Bayesian vignettes to be more semantically clear and concise (e.g., "strong evidence for no adverse impact by gender" or "strong evidence against adverse impact," rather than "strong evidence supporting that hiring rates do not differ by gender").

Moreover, "weakness" is more semantically congruent with a lack of evidence, whereas the use of the word "strong" may have been confused in this context. This was the only vignette with a lack of congruency in interpretation, because although it is possible, it is generally implausible to have a very high p value associated with such a large effect size. Scenarios with evidence supporting adverse impact used more congruent language (i.e., "very large effect size" and "very strong evidence") because the direction of the effects leads to congruent language. Frequentist testing is unipolar, going from rejection to a failure to reject the null hypothesis, whereas Bayes factors are bipolar, ranging from very strong evidence supporting the null hypothesis to very strong evidence supporting the alternative hypothesis.

In sum, the word "weak" may have been more synonymous with the violation of the 4/5ths Rule for participants with lower levels of statistical knowledge. Wording differences (i.e., a very weak effect versus strong evidence supporting that hiring rates do not differ by gender), in conjunction with participants completing the study faster than expected based on piloting, may have also contributed to the disparity in interpretability ratings.

Limitations and Future Directions

The limitations of our study design and study sample can provide inspiration for future research. For example, we asked participants to read two vignettes while imagining they were jury members, but future research might attempt to make the jury setting more realistic within a quasi-experimental design (e.g., survey those who have had or will have jury experiences related to employment discrimination). As another example, the vignettes contained succinct pieces of information that, in an actual jury setting, would be included in a more extensive trial process. A more elaborate follow-up experiment could provide evidence and statements from both the prosecution and defense, along with a group decision-making process, and perhaps decisions about adverse impact could be made multiple times by participants during this mock-jury process. Furthermore, it would be interesting to assess interpretability in a study in which the results are spoken rather than written, as this could increase the external validity of the results with respect to courtroom situations.

Another future extension of our study design would be to manipulate the 4/5ths Rule violations alongside related statistical evidence. In our vignettes, we used the same extreme impact ratios across Bayesian and frequentist paradigms (i.e., very strong evidence of adverse impact vs. no adverse impact), only varying the strength of statistical evidence supporting those ratios. A future study could vary the impact ratio alongside the strength of statistical evidence (e.g., moderate or small evidence of adverse impact vs. no adverse impact). This would serve to refine and generalize our understanding further, with regard to the relationship between participant perceptions of the interpretability and retributive justice of adverse impact vignettes.

It is possible that we observed no difference between paradigms on most measures because participants were making decisions about adverse impact based solely on the 4/5ths Rule without considering accompanying statistical evidence. This limitation might imply a lesson in adverse impact decision-making: Raters should be further trained to understand and consider statistical evidence in support of the impact ratio and not just solely focus on the latter. For example, participants could be trained in both Bayesian and frequentist paradigms prior to completing studies similar to ours. This could serve to control for prior statistical knowledge, and generalize to real-world settings where such training is beneficial. In the current study, the lack of knowledge and consequential inability to interpret statistical evidence may together have contributed to participants completing the study faster than expected, because they may not have been reading the vignettes in their entirety. Nonetheless, our study aimed to be realistic, being based on samples with naturally lower levels of statistical knowledge and no statistical training given, who are reading vignettes that reflect how Bayesian and frequentist information is typically communicated.

Future studies could also examine the presence (and interpretability) of adverse impact using posterior distributions of the impact ratio, rather than the chi-square test. As we stated with the Chandler (2020) work, Bayesian analyses may be more interpretable for parameter estimation than for hypothesis testing. This comparison could be examined with evidence provided in adverse impact cases as well. In sum, although we were unable to support our hypotheses, our study still provides evidence that

Bayesian results are at least as interpretable as frequentist results regarding perceptions of adverse impact, thus providing initial promise for their use, alongside the general advantages of Bayesian analyses that were reviewed in our introduction. We hope that future research will build further on this promise.

Appendix 1

Vignette 1

Imagine that you are selected to serve on a jury during the trial of Smith v. Bright Light. Bright Light Company, which manufactures and sells lightbulbs, is on trial for adverse impact after a recent job candidate, Ms. Smith, claimed that she was not hired because of her gender. Ms. Smith's lawyer hired a subject matter expert, Dr. Williams, to present statistical evidence to you and the other jurors at the trial.

The prosecution calls Dr. Williams to the stand. In response to the prosecution's question about statistical evidence, Dr. Williams reports (a) the raw data, (b) results from applying the 4/5ths Rule, and (c) results from a chi-square test of independence as follows:

"A total of 60 people applied (40 men, 20 women) and 25 people were hired (23 men, 2 women)."

"In addition, the 4/5ths Rule indicates evidence of adverse impact, because 10% of the women and 57.5% of the men were hired, resulting in a ratio of .17 which is less than .80 (or 4/5ths)."

"A chi-square test was used to determine if the observed hiring rate of men and women differed from a fair hiring rate (i.e., where the hiring rate is completely independent of gender). We observed the hiring rate of men compared to women being statistically significant (p = .002), with men being hired at a greater rate than women; thus we can reject the null hypothesis that the hiring rate is independent from gender. This corresponds to a phi-coefficient of .39, which by conventional rules of thumb constitutes a very large effect size."

Vignette 2

Imagine that you are selected to serve on a jury during the trial of Smith v. Bright Light. Bright Light Company, which manufactures and sells lightbulbs, is on trial for adverse impact after a recent job candidate, Ms. Smith, claimed that she was not hired because of her gender. Ms. Smith's lawyer hired a subject matter expert, Dr. Williams, to present statistical evidence to you and the other jurors at the trial.

The prosecution calls Dr. Williams to the stand. In response to the prosecution's question about statistical evidence, Dr. Williams reports (a) the raw data, (b) results from applying the 4/5ths Rule, and (c) results from a chi-square test of independence as follows:

"A total of 60 people applied (40 men, 20 women) and 25 people were hired (23 men, 2 women)."

"In addition, the 4/5ths Rule indicates evidence of adverse impact because 10% of the women and 57.5% of the men were hired, resulting in a ratio of .17 which is less than .80 (or 4/5ths)."

"A chi-square test was used to determine if the observed hiring rate of men and women differed from a fair hiring rate (i.e., where the hiring rate is completely independent of gender). The alternative hypothesis, that the hiring rate is dependent on gender, is 38 times better at explaining the data than the null hypothesis that the hiring rate is independent from gender ($BF_{10} = 38$). By conventional rules of thumb, this constitutes very strong evidence favoring that men are hired at a greater rate than women."

Vignette 3

Imagine that you are selected to serve on a jury during the trial of Johnson v. All Your Appliances Company, which manufactures and sells appliances, is on trial for adverse impact after a recent job candidate, Ms. Johnson, claimed that she was not hired because of her gender. All Your Appliances Company's lawyer hired a subject matter expert, Dr. Jones, to present statistical evidence to you and the other jurors at the trial.

The defense calls Dr. Jones to the stand. In response to the defense's question about statistical evidence, Dr. Jones reports (a) the raw data, (b) results from applying the 4/5ths Rule, and (c) results from a chi-square test of independence as follows:

"A total of 315 people applied (157 men, 158 women) and 4 people were hired (2 men, 2 women)."

"In addition, the 4/5ths Rule does not indicate evidence of adverse impact because 1.3% of women are hired and 1.3% of men are hired, resulting in a ratio of .99 which is greater than .80 (or 4/5ths)."

"A chi-square test was used to determine if the observed hiring rate of men and women differed from a fair hiring rate (i.e., where the hiring rate is completely independent of gender). There is not enough evidence to determine if there is a difference in hiring rates based on gender (p = .995). This corresponds to a phi-coefficient of .0003, which is classified as a very weak effect."

Vignette 4

Imagine that you are selected to serve on a jury during the trial of Johnson v. All Your Appliances Company, which manufactures and sells appliances, is on trial for adverse impact after a recent job candidate, Ms. Johnson, claimed that she was not hired because of her gender. All Your Appliances Company's lawyer hired a subject matter expert, Dr. Jones, to present statistical evidence to you and the other jurors at the trial.

The defense calls Dr. Jones to the stand. In response to the defense's question about statistical evidence, Dr. Jones reports (a) the raw data, (b) results from applying the 4/5ths Rule, and (c) results from a chi-square test of independence as follows:

"A total of 315 people applied (157 men, 158 women) and 4 people were hired (2 men, 2 women)."

"In addition, the 4/5ths Rule does not indicate evidence of adverse impact because 1.3% of women are hired and 1.3% of men are hired, resulting in a ratio of .99 which is greater than .80 (or 4/5ths)."

"A chi-square test was used to determine if the observed hiring rate of men and women differed from a fair hiring rate (i.e., where the hiring rate is completely independent of gender). The null hypothesis that the rate of females hired is independent of the rate of males hired is 30 times better at explaining the data than the alternative hypothesis that hiring is dependent on gender. By conventional rules of thumb, this constitutes strong evidence supporting that hiring rates do not differ by gender."

Appendix 2

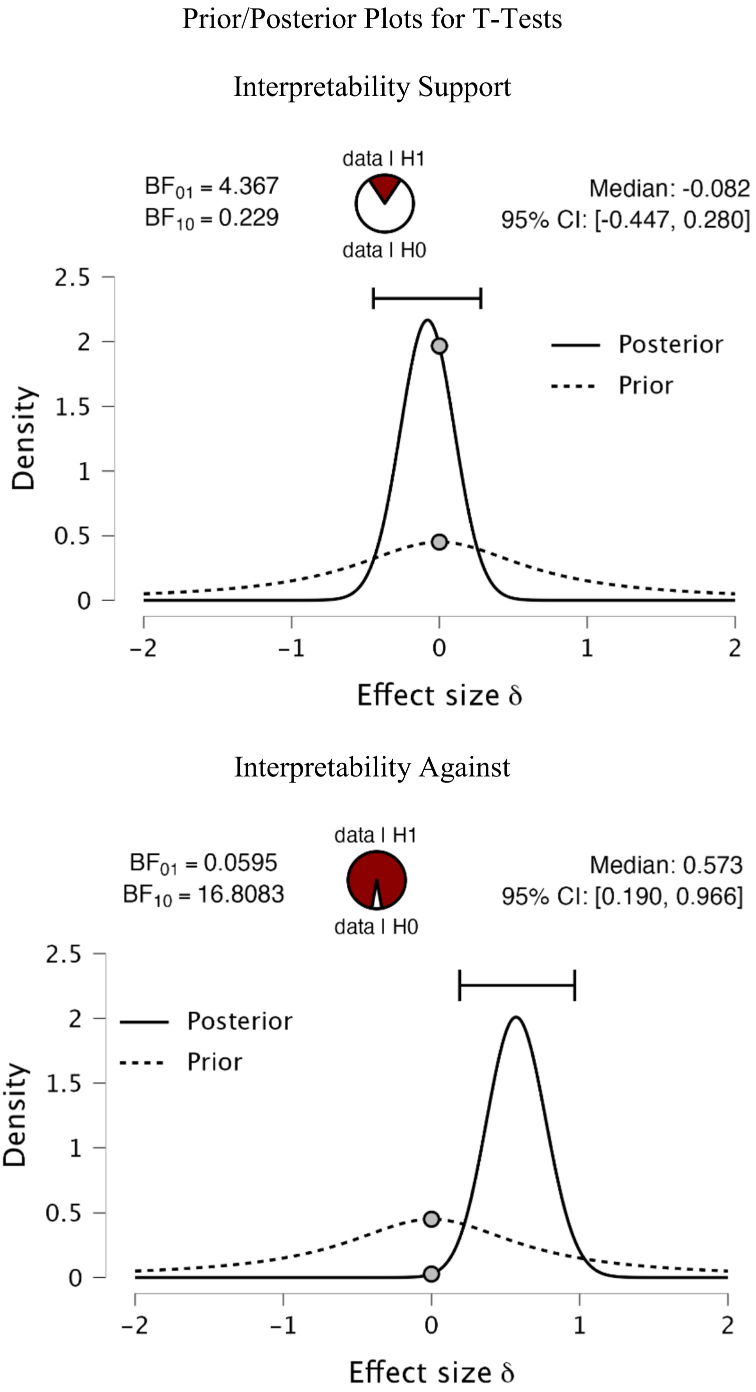
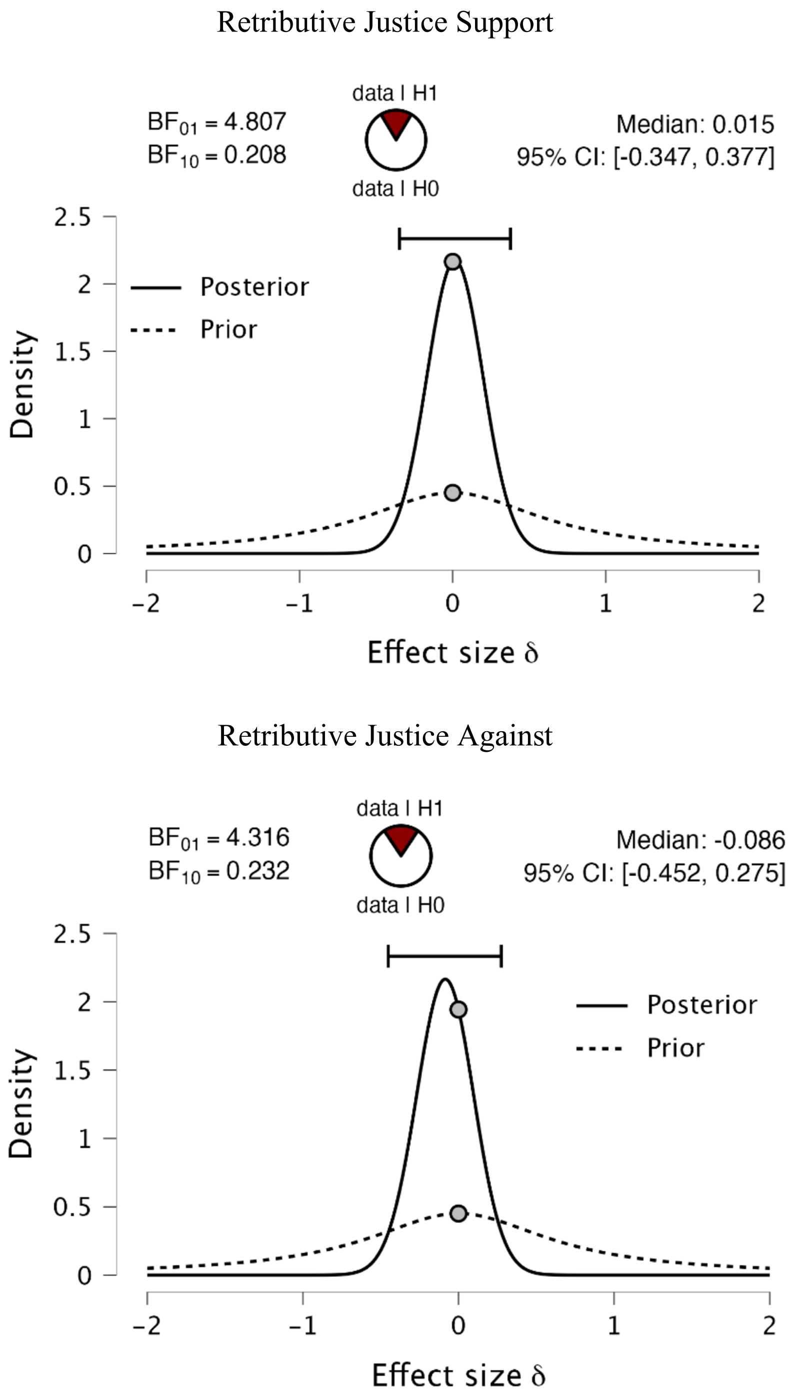


Fig. 6 Prior/posterior plots for *t*-tests

Fig. 6 (continued)



Retributive Justice Support

Retributive Justice Against

Appendix 3

Fig. 7 Bayes factor robustness checks for *t*-tests

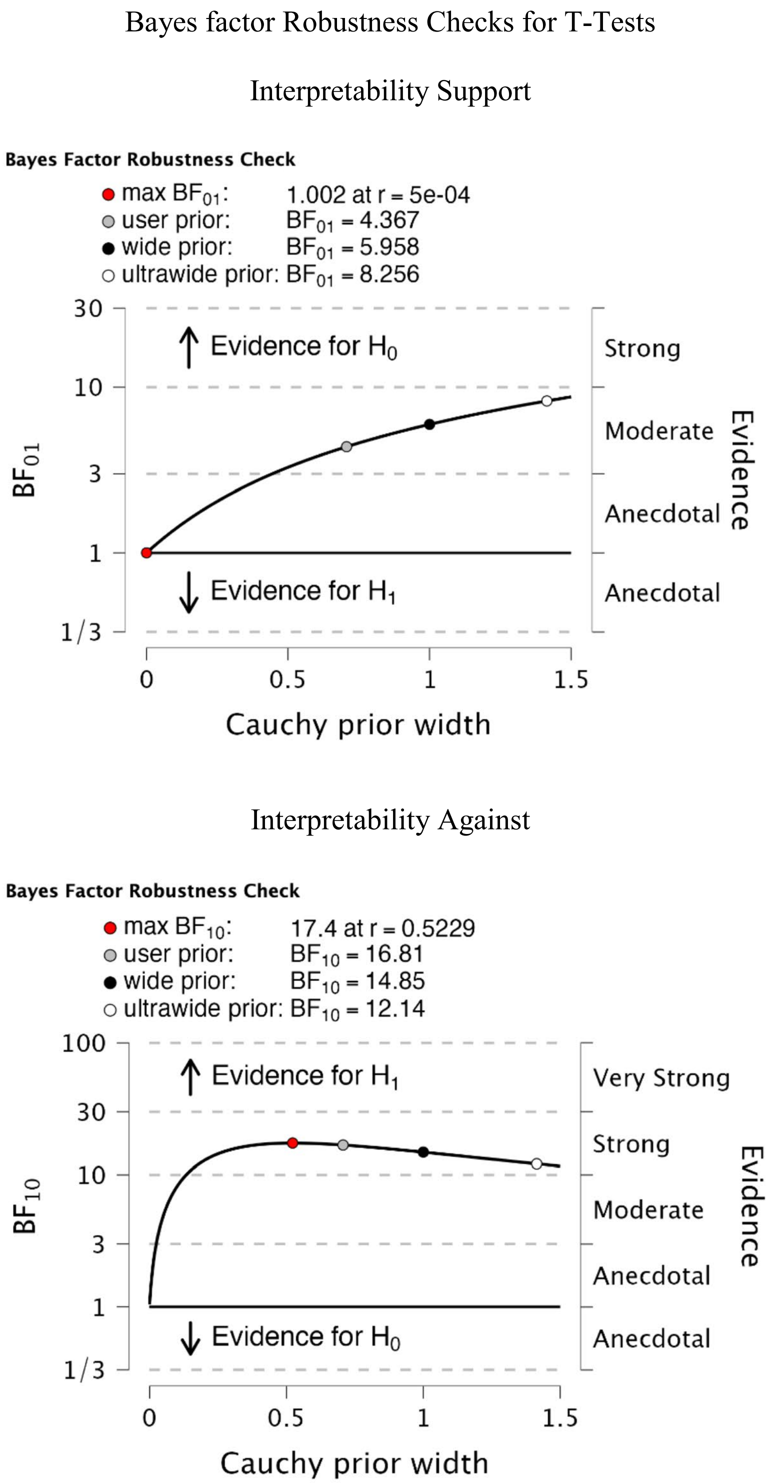
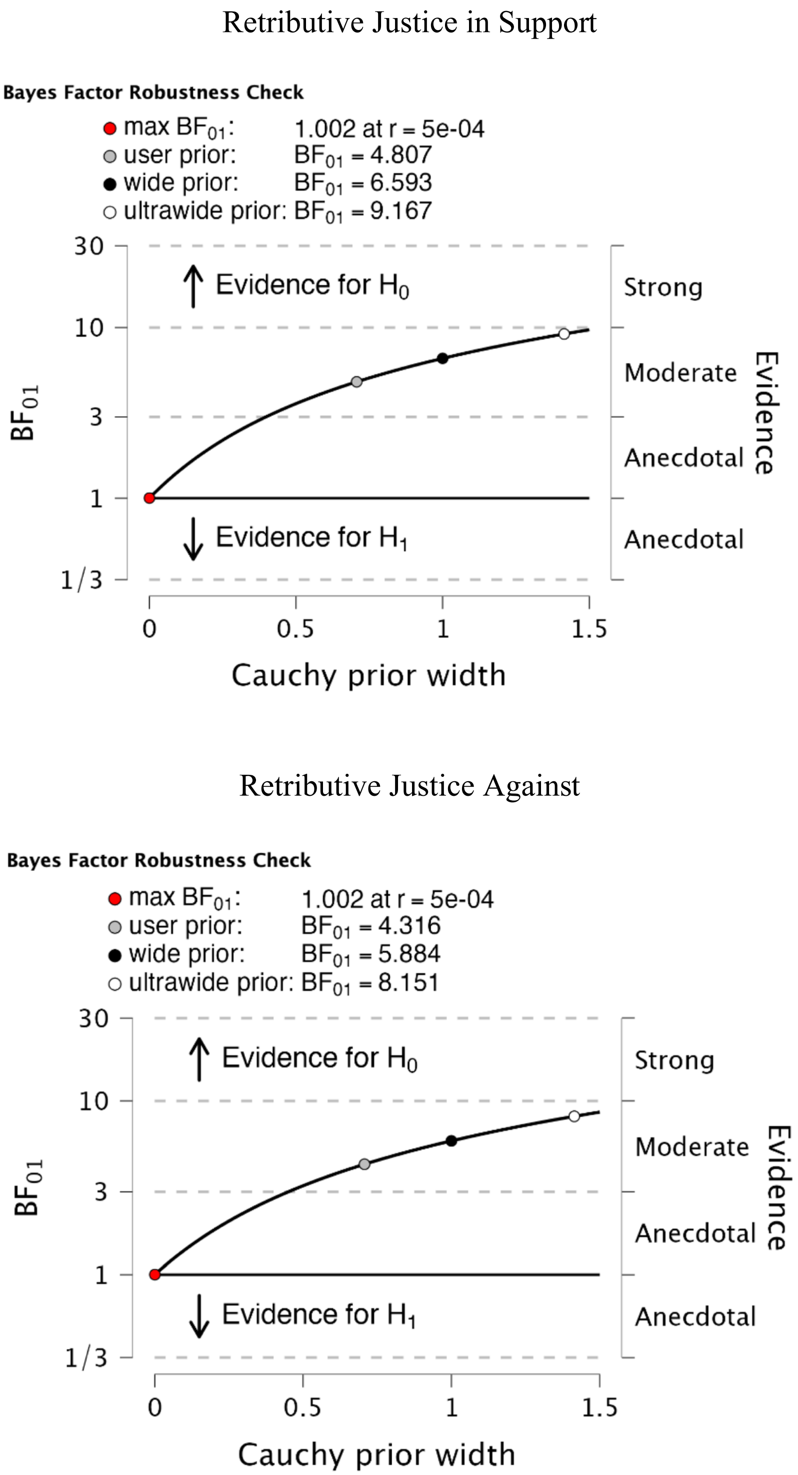


Fig. 7 (continued)



Appendix 4

Bayes factor model comparison to best model

Table 7 Model comparison relative to the best model: Interpretability and retributive justice

	Interpretability		Retributive justice
	BF_{01} (Best vs. Others)		BF_{01} (Best vs. Others)
Null model (Grand mean)	11.42	Null model (Grand mean)	> 100
Best model (Measure + Paradigm + Measure*Paradigm)	1.00	Best model (Measure)	1.00
Measure	14.01 [a]	Measure + Paradigm	5.72
Measure + Paradigm	17.19	Measure + Paradigm + Measure*Paradigm	26.60
Paradigm	13.55	Paradigm	> 100

$N = 104$ ($n = 54$ frequentist; $n = 50$ Bayesian). BF_{01} = Bayes factor compared to best model (other = 1, best = 0); numbers greater than 1 indicate support for the best model and numbers less than 1 indicate support for the other model. All error percentages were less than 20% (see van Doorn, 2021)

[a] For example, the best interpretability model Best model (Measure + Paradigm + Measure*Paradigm) was 14.01 times better at explaining the data than the Measure model

Declarations

Conflict of Interest The authors declare no competing interests.

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., III., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45*(6), 721–734. https://doi.org/10.1037/0003-066X.45.6.721

Ballard, T., Vancouver, J. B., & Neal, A. (2018). On the pursuit of multiple goals with different deadlines. *Journal of Applied Psychology, 103*(11), 1242–1264. https://doi.org/10.1037/apl0000304

Bobko, P., & Roth, P. L. (2004). The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In J. J. Martocchio (Ed.), *Research in personnel and human resources management*. Elsevier Science/JAI Press. 23, 177–198 https://doi.org/10.1016/S0742-7301(04)23004-3

Chandler, J. J., Martinez, I., Finucane, M. M., Terziev, J. G., & Resch, A. M. (2020). Speaking on data's behalf: What researchers say and how audiences choose. *Evaluation Review, 44*(4), 325–353. https://doi.org/10.1177/0193841X19834968

Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000e et seq (1964).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*, 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Goodman, S. (2008). A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology, 48*(4), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Grand, J. A. (2017). Brain drain? An examination of stereotype threat effects during training on knowledge acquisition and organizational effectiveness. *Journal of Applied Psychology, 102*(2), 115–150. https://doi.org/10.1037/apl0000171

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I Foundations and basic methods. *International Journal of Epidemiology, 35*(3), 765–775. https://doi.org/10.1093/ije/dyi312

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. https://doi.org/10.1007/s10654-016-0149-3

Gronau, Q. F., Ly, A., & Wagenmakers, E. J. (2019). Informed Bayesian *t*-tests. *The American Statistician, 74*, 137–143. https://doi.org/10.1080/00031305.2018.1562983

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p*-values. *Psychonomic Bulletin & Review, 13*, 1033–1037. https://doi.org/10.3758/BF03213921

Hurwitz, A. (2020). Is the glass half empty or half full?: An experimental study of Bayesian versus frequentist statistics' influence on program endorsements by legislative staff [Thesis]. https://udspace.udel.edu/handle/19716/28565

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Kim, Y.-J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101*(7), 976–994. https://doi.org/10.1037/apl0000102

Jebb, A. T., & Woo, S. E. (2015). A Bayesian primer for the organizational sciences: The "two sources" and an introduction to BugsXLA. *Organizational Research Methods, 18*(1), 92–132. https://doi.org/10.1177/1094428114553060

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Kline, R. B. (2020). Post *p* value education in graduate statistics: Preparing tomorrow's psychology researchers for a postcrisis future. *Canadian Psychology/psychologie Canadienne, 61*(4), 331–341. https://doi.org/10.1037/cap0000200

Kruschke, J. K. (2010). Bayesian data analysis. *WIREs. Cognitive Science, 1*(5), 658–676. https://doi.org/10.1002/wcs.72

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139087759

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

Morey, R. D., & Rouder, J. N. (2021). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12–4.3. https://CRAN.Rproject.org/package=BayesFactor

Morris, S. B. (2001). Sample size required for adverse impact analysis. *Applied HRM Research, 6*(1–2), 13–32.

Morris, S., & Lobsenz, R. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*(1), 89–111. https://doi.org/10.1111/j.1744-6570.2000.tb00195.x

Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-analysis. *Journal of Applied Psychology, 92*(5), 1394–1413. https://doi.org/10.1037/0021-9010.92.5.1394

Oswald, F. L., Wu, F. Y., & Courey, K. A. (2021). Training (and retraining) in data, methods, and theory in the organizational sciences. In K. R. Murphy (Ed.), *data, methods, and theory in the organizational sciences* (pp. 294–316). Routledge.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Revelle, W. (2016) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psychVersion=1.6.4.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Uniform Guidelines on Employee Selection Procedures (UGESP). (1978). 43 Fed. *Reg., 38295*, 38290–38315.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., ... & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1-26.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., ... & Wagenmakers, E. J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813-826. https://doi.org/10.3758/s13423-020-01798-5

van Prooijen, J. W., & Coffeng, J. (2013). What is fair punishment for Alex or Ahmed? Perspective taking increases racial bias in retributive justice judgments. *Social Justice Research, 26*(4), 383–399. https://doi.org/10.1007/s11211-013-0190-2

van Ravenzwaaij, D., & Etz, A. (2021). Simulation studies as a tool to understand Bayes factors. *Advances in Methods and Practices in Psychological Science, 4*, 1–20. https://doi.org/10.1177/2515245920972624

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Walen, A. (2021). Retributive justice. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wenzel, M., & Okimoto, T. G. (2016). Retributive justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of social justice theory and research* (pp. 237–256). Springer. https://doi.org/10.1007/978-1-4939-3216-0

Wong, T. K., Kiers, H., & Tendeiro, J. (2021). *On the potential mismatch between the function of the Bayes factor and researchers' expectations*. PsyArXiv.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41*(2), 390–420. https://doi.org/10.1177/0149206313501200

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.