METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Tutorial

Check for updates

Bayesian Tests of Two Proportions: A Tutorial With R and JASP

Tabea Hoffmann [1], Abe Hofman [2], Eric-Jan Wagenmakers [2]

[1] *Faculty of Economics and Business & Faculty of Spatial Sciences, University of Groningen, Groningen, The Netherlands.* **[2]** *Psychological Methods Group, University of Amsterdam, Amsterdam, The Netherlands.*

Methodology, 2022, Vol. 18(4), 239–277, https://doi.org/10.5964/meth.9263

Received: 2022-04-04 • **Accepted:** 2022-09-05 • **Published (VoR):** 2022-12-22

Handling Editor: Isabel Benítez, University of Granada, Granada, Spain

Corresponding Author: Tabea Hoffmann, Faculty of Economics and Business & Faculty of Spatial Sciences, Nettelbosje 2, University of Groningen, L9747 AE Groningen, The Netherlands. E-mail: t.hoffmann@rug.nl

Supplementary Materials: Data [see Index of Supplementary Materials]

Abstract

The need for a comparison between two proportions (sometimes called an A/B test) often arises in business, psychology, and the analysis of clinical trial data. Here we discuss two Bayesian A/B tests that allow users to monitor the uncertainty about a difference in two proportions as data accumulate over time. We emphasize the advantage of assigning a dependent prior distribution to the proportions (i.e., assigning a prior to the log odds ratio). This dependent-prior approach has been implemented in the open-source statistical software programs R and JASP. Several examples demonstrate how JASP can be used to apply this Bayesian test and interpret the results.

Keywords

Bayes factor, Bayesian estimation, contingency tables, log odds ratio

The comparison of two sample proportions is also known as an A/B test (Little, 1989). The statistical framework requires that there are two groups and that the data for each participant are dichotomous, e.g., 'correct–incorrect', 'infected–not infected', or 'watched the commercial–did not watch the commercial'.

The A/B test for proportions (henceforth 'the A/B test') is standard operating procedure for the analysis of clinical trial data: study participants are randomly allocated to one of two experimental conditions (which are often called group A and group B). One of the conditions is usually the control condition (e.g., a placebo), while the other condi-

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.

tion introduces an intervention (e.g., a new pharmacological drug). In each condition, participants are evaluated on a dichotomous measure such as 'dead–alive', 'side effect–no side effect', etc. The goal of the experiment is to examine the treatment success of the intervention. Because of its general nature, the A/B test is also common in fields such as biology and psychology.

Another field that has recently adopted A/B testing—for so-called 'conversion rate optimization'—is online marketing. A conversion rate optimization experiment proceeds analogously to the classical experiment: two versions of the same website are shown to different selections of website visitors and the number of visitors who take a desired action (e.g., clicking on a specific button) is monitored.

Whenever the A/B test is applied, practitioners eventually wish to know whether and to what extent the experimental condition has a higher success rate than the control condition. This judgment requires that the observed sample difference in proportions is translated to the population—that is, the judgment requires statistical inference.

In general, practitioners can choose between the frequentist and the Bayesian frameworks for their statistical analysis. We subscribe to the three general desiderata for inference in the A/B test as outlined in Gronau, Raj K. N., and Wagenmakers (2021): ideally, (1) evidence can be obtained in favor of the null hypothesis; (2) evidence can be monitored continually, as the data accumulate; and (3) expert knowledge can be taken into account. Below we briefly explain why these desiderata are incompatible with the framework of frequentist statistics; next we turn to the Bayesian framework and examine two different Bayesian instantiations of the A/B test, which we then apply to two examples. This article is a tutorial, and consequently we will emphasize the assumptions, interpretations, and practical application of the test. A more advanced discussion of the software can be found in Gronau et al. (2021), and an associated statistical paradox is presented in Dablander, Huth, Gronau, Etz, and Wagenmakers (2022).

Frequentist Statistics

Practitioners predominantly use *p*-value null-hypothesis significance testing (NHST) to analyze A/B test data. However, the standard NHST approach does not satisfy the three desiderata mentioned above. Firstly, standard NHST results cannot distinguish between *absence of evidence and evidence of absence* (Keysers, Gazzola, & Wagenmakers, 2020; Robinson, 2019). Evidence of absence means that the data support the hypothesis that there is no effect (i.e., the two conditions do not differ); absence of evidence, however, means that the data are inconclusive (Altman & Bland, 1995). Secondly, in standard NHST the data cannot be tested sequentially without necessitating a correction for multiple comparisons that depends on the sampling plan (see for instance Berger & Wolpert, 1988; Wagenmakers, 2007; Wagenmakers et al., 2018). Especially in clinical trials but also for online marketing is it efficient to act as soon as the data provide evidence that is sufficiently compelling. To achieve such efficiency many A/B test practitioners

PsychOpen GOLD

repeatedly peek at interim results and stop data collection as soon as the p-value is smaller than some predefined α-level (Goodson, 2014). However, this practice inflates the Type I error rate and hence invalidates an NHST analysis (Jennison & Turnbull, 1990; Wagenmakers, 2007). Thirdly, standard NHST does not allow users to incorporate detailed expert knowledge. For example, among conversion rate optimization professionals it is widely known that online advertising campaigns often yield minuscule increases in conversion rates (cf. Johnson, Lewis, & Nubbemeyer, 2017; Patel, 2018). Such knowledge may affect NHST planning (i.e., knowledge that the effect is minuscule would necessitate the use of very large sample sizes), but it is unclear how it would affect inference.[1] As we will see below, in the Bayesian framework is it conceptually straightforward to enrich statistical models with expert background knowledge, thereby resulting in more informed statistical analyses (Lindley, 1993).

It should be acknowledged, however, that non-standard (i.e., less popular) forms of frequentist analyses exist that alleviate some of the concerns listed above. For instance, sequential inference can be carried out by the Sequential Probability Ratio Test (Schnuerch & Erdfelder, 2020; Wald, 1945) or by Safe Testing (e.g., Grünwald, de Heide, & Koolen, 2021). In addition, it has been argued that evidence of absence can be obtained by means of an equivalence test, in which the null hypothesis is defined as an effect size that falls outside of a region of practical interest (e.g., King, 2011; Tango, 1998). An in-depth discussion of the pros and cons of frequentist inference is beyond the scope of this article.

Bayesian Statistics

The limitations of standard frequentist statistics can be overcome by adopting a Bayesian data analysis approach (e.g., Deng, 2015; Kamalbasha & Eugster, 2021; Stucchio, 2015). In Bayesian statistics, probability expresses a degree of knowledge or reasonable belief (Jeffreys, 1961) and in principle Bayesian statistics fulfills all three desiderata listed above (e.g., Wagenmakers et al., 2018). In the next sections we introduce two approaches to Bayesian A/B testing. The two approaches make different assumptions, ask different questions, and therefore provide different answers (cf. Dablander et al., 2022).

The 'Independent Beta Estimation (IBE) Approach'

Let n_A denote the total number of observations and y_A denote the number of successes for Group A. Let n_B denote the total number of observations and y_B denote the number

1) For example, with the data in hand one may find that $p = 0.15$, and that the power to detect a minuscule effect was only 0.20. However, power is a pre-data concept and consequently it remains unclear to what extent the observed data affect our knowledge (Wagenmakers et al., 2015). Moreover, the selection of the minuscule effect is often motivated by Bayesian considerations (i.e., it is a value that appears plausible, based on substantive domain knowledge).

PsychOpen GOLD

of successes for Group B. The commonly used Bayesian A/B testing model is specified as follows:

$$y_A \sim Binomial(n_A, \theta_A)$$
$$y_B \sim Binomial(n_B, \theta_B)$$

This model assumes that y_A and y_B follow independent binomial distributions with success probabilities θ_A and θ_B. These success probabilities are assigned independent beta(α, β) distributions that encode the relative prior plausibility of the values for θ_A and θ_B. In a beta distribution, the α values can be interpreted as counts of hypothetical 'prior successes' and the β values can be interpreted as counts of hypothetical 'prior failures' (Lee & Wagenmakers, 2013):

$$\theta_A \sim Beta(\alpha_A, \beta_A)$$
$$\theta_B \sim Beta(\alpha_B, \beta_B)$$

Data from the A/B testing experiment update the two independent prior distributions to two independent posterior distributions as dictated by Bayes' rule:

$$p(\theta_A \mid y_A, n_A) = \frac{p(\theta_A) \times p(y_A, n_A \mid \theta_A)}{p(y_A, n_A)}$$
$$p(\theta_B \mid y_B, n_B) = \frac{p(\theta_B) \times p(y_B, n_B \mid \theta_B)}{p(y_B, n_B)}$$

where $p(\theta_A)$ and $p(\theta_B)$ are the prior distributions and $p(y_A,n_A \mid \theta_A)$ and $p(y_B,n_B \mid \theta_B)$ are the likelihoods of the data given the respective parameters. Hence, the reallocation of probability from prior to posterior is brought about by the data: the probability increases for parameter values that predict the data well and decreases for parameter values that predict the data poorly (Kruschke, 2013; van Doorn, Matzke, & Wagenmakers, 2020; Wagenmakers, Morey, & Lee, 2016). Note that whenever a beta prior is used and the observed data are binomially distributed, the resulting posterior distribution is also a beta distribution. Specifically, if the data consist of s successes and f failures, the resulting posterior beta distribution equals beta($\alpha + s$, $\beta + f$) (Gelman et al., 2013; van Doorn et al., 2020).[2] Ultimately, practitioners are most often interested in the difference $\delta = \theta_A - \theta_B$ between the success rates of the two experimental groups, as this difference indicates whether the experimental condition shows the desired effect.

2) When the prior and the posterior belong to the same family of distributions they are said to be *conjugate*.

PsychOpen GOLD

R Implementation of the IBE Approach: The `bayesAB` Package — The IBE approach is implemented for instance in the `bayesAB` (version 1.1.3, Portman, 2017) package in R (version 4.2.1, R Core Team, 2020). Consider the following fictitious example from ethology, inspired by the classic work of von Frisch (1914). A researcher wishes to test whether honey bees have color vision by comparing the behavior of two groups of bees. The experiment involves a training and a testing phase. In the training phase, the bees in the experimental condition are presented with a blue and a green disc. Only the blue disc is covered with a sugar solution that bees crave. The control group receives no training. In the testing phase, the sugar solution is removed from the blue disc, and the behavior of both groups is being observed. If the bees in the experimental condition have learned that only the blue disc contains the appetising sugar solution, and if they can discriminate between blue and green, they should preferentially explore the blue instead of the green disc during the testing phase. The researcher finds that in 65 out of 100 times, the bees in the experimental group continued to approach the blue disc after the sugar solution was removed. The bees that were not trained approached the blue disc 50 out of 100 times. In the remainder of this section, we will refer to the bees in the control condition as group A and to the bees in the experimental condition as group B. The R file for this fictitious example can be found in the Supplementary Materials.

Before setting up this A/B test and collecting the data, the prior distribution has to be specified so that it represents the relative plausibility of the parameter values. For the present example, the researcher specifies two uninformative (uniform) beta(1,1) priors. After running the A/B test procedure, the priors are updated with the obtained data. With `bayesAB` the calculation of the posterior distributions is done by feeding both the priors and the data to the `bayesTest` function:

```
R> library(bayesAB)
R> bees1 <- read.csv2("bees_data1.csv")
R> AB1 <- bayesTest(bees1$y1, bees1$y2,
+                   priors = c('alpha' = 1, 'beta' = 1),
+                   n_samples = 1e5, distribution = 'bernoulli')
```

A more detailed explanation of the function and its arguments can be obtained by typing `?bayesTest` into the R console. The results can be obtained and visualized by executing:

```
R> summary(AB1)
R> plot(AB1)
```

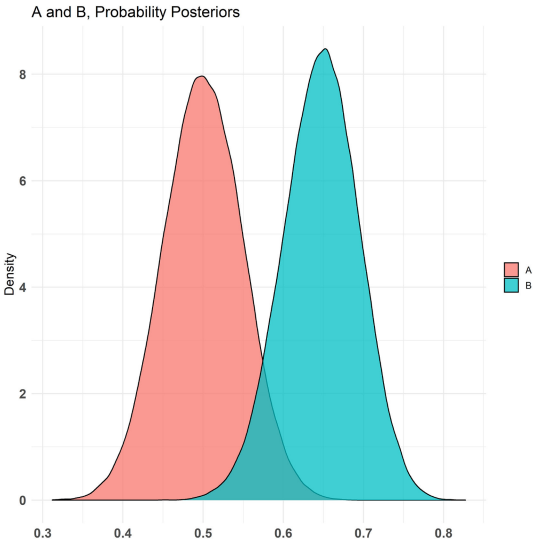
Figure 1 shows the two independent posterior distributions that `plot(AB1)` returns. To plot these posterior distributions, `bayesTest` makes use of the `rbeta` function that draws random numbers from a given beta distribution. To obtain each posterior distribution the package first exploits conjugacy: the number of successes s are added to the α values of either version's prior distribution and the number of failures f are

PsychOpen GOLD

added to the respective β values (e.g., Kruschke, 2015; Kurt, 2019). Thus, the posterior distribution for θ_A is beta($\alpha_A + s_A$, $\beta_A + f_A$) and that for θ_B is beta($\alpha_B + s_B$, $\beta_B + f_B$). The `rbeta` function draws random samples from each posterior distribution and the density of these values is shown in Figure 1.[3] We can see that group B's posterior distribution for the success probability assigns more mass to higher values of θ. This suggests that the success probability of the trained bees is higher, which in turn implies that bees have color vision.

Figure 1

Independent Posterior Beta Distributions of the Success Probabilities for Group A and B



Note. The plot is produced by the `bayesAB` package with the fictitious bee data (i.e., A = 50/100 versus B = 65/100) described in the main text. The analysis used two independent beta(1, 1) priors.

The `bayesAB` package also returns a posterior distribution which indicates the 'conversion rate uplift' $\frac{(\theta_B - \theta_A)}{\theta_A}$, that is, the difference between the success rates expressed as a proportion of θ_A. The advantage of expressing the difference as a proportion of θ_A is that a change from 1% to 2% (i.e., a doubling of the conversion rate) is seen to be much more impressive than a change from 50% to 51%. The associated disadvantage is that a small change can appear more impressive than it really is. The posterior distribution for the conversion rate uplift is computed from the random samples obtained for the two beta

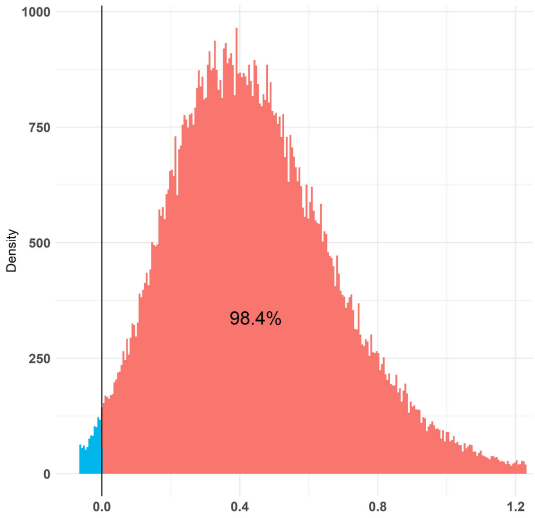
3) The posterior distributions are available analytically, so at this point the rbeta function is not needed; it will become relevant once we start to investigate the posterior distribution for the difference between θ_A and θ_B.

PsychOpen GOLD

posteriors shown in Figure 1. As shown in Figure 2, the posterior distribution for the uplift peaks at around 0.4, indicating that the most likely increase of bee approaches on the blue disc equals 40%. Also, most posterior mass (i.e., 98.4% of the samples) is above zero, indicating that we can be 98.4% certain that group B approaches the blue disc more often than group A. Note that this statement assumes that it is a priori equally likely that the training in the experimental condition B had a positive or negative effect on the rate at which the bees approach the blue disc, and that the possibility that both groups have the same approach rate is deemed impossible from the outset—this is an important point to which we will return later.

Figure 2

Histogram of the Conversion Rate Uplift From Version A (i.e., 50/100) to Version B (i.e., 65/100) for the Fictitious Bee Data Set



Note. The uplift is calculated by dividing the difference in conversion by the conversion in A. The plot is produced by the `bayesAB` package.

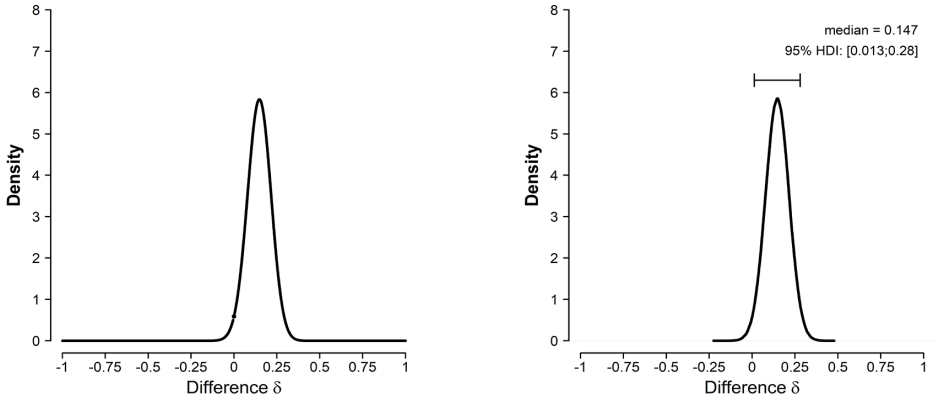
The posterior probability that $\theta_B > \theta_A$ can also be obtained analytically (Schmidt & Mørup, 2019). The formula is not implemented in the `bayesAB` package; our R implementation can be found in the OSF repository (Hoffmann, Hofman, & Wagenmakers, 2022). For the above example, $p(\theta_B > \theta_A \mid \text{data}) = 0.984$. In fact the entire posterior distribution for the difference between the two independent beta distributions is available analytically (Pham-Gia, Turkkan, & Eng, 1993). The left-hand panel of Figure 3 shows the posterior distribution of the difference between the two independent beta posteriors from the bee example. Unfortunately, the analytic calculation fails for values of α and β above ~ 70, which occur with strong advance knowledge or high sample sizes.[4] In this

PsychOpen GOLD

case, one can instead employ a normal approximation, the result of which is shown in the right panel of Figure 3.[5]

Figure 3

Posterior Distributions of the Difference $\delta = \theta_B - \theta_A$ for the Fictitious Bee Data (i.e., A = 50/100 Versus B = 65/100)



Note. The left-hand panel shows the analytic distribution of the difference between two independent beta distributions (Pham-Gia et al., 1993). The right-hand panel shows the normal approximation of the difference between two independent beta distributions.

One advantage of the Bayesian approach is that the data can also be added to the analysis in a sequential manner. This means that the evidence can be assessed continually as the data arrives and the analyses can be stopped as soon as the evidence is judged to be compelling (Deng, Lu, & Chen, 2016). As a demonstration, Figure 4 plots the posterior mean of the difference between θ_A and θ_B as well as the 95% highest density interval (HDI) of the difference in a sequential manner. The HDI narrows with increasing sample size, indicating that the range of likely values for δ gradually becomes smaller. After some initial fluctuation, the posterior mean difference between θ_A and θ_B (i.e., the orange line) settles between 0.1 and 0.2. The R code for the sequential computation can be found in the OSF repository (Hoffmann, Hofman, & Wagenmakers, 2022). Note that the results are not analytic—they are based on repeatedly drawing samples. This sampling process introduces variability such that when the same analysis is executed again on the same data, the outcomes will differ slightly. The numerical variability can be made arbitrarily small by drawing more samples.

4) The reason for this is numerical overflow from Appell's first hypergeometric function.

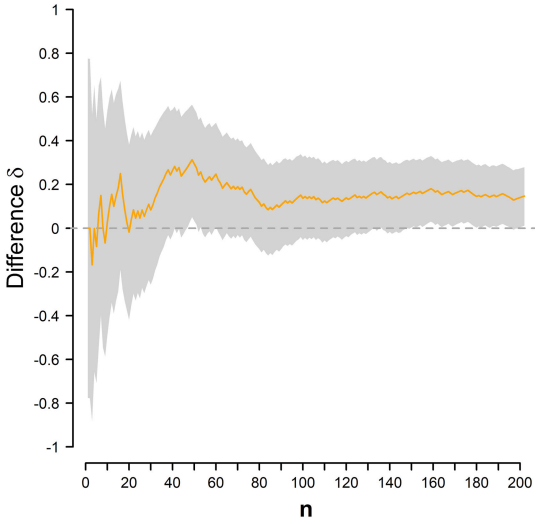
5) The appendix contains the formulas from Schmidt and Mørup (2019) and Pham-Gia et al. (1993), as well as the formulas for the normal approximation.

PsychOpen GOLD

In sum, the IBE approach allows practitioners to judge the size and direction of an effect, that is, the difference between the two success probabilities. It is important, however, to recognize the assumptions that come with this approach. In the next section, we will elaborate on these assumptions and their consequences.

Figure 4

Sequential Analysis of the Difference Between the Success Probabilities (i.e., $\delta = \theta_B - \theta_A$) for the Fictitious Bee Data (i.e., A = 50/100 versus B = 65/100)



Note. The orange line represents the posterior mean of δ. The grey area visualizes the width of the 95% HDI as a function of sample size *n*.

Assumptions of the IBE Approach — The IBE approach makes two important assumptions. The first assumption is that the two success probabilities are independent: learning about the success rate of one experimental condition does not affect our knowledge about the success rate of the other condition (Howard, 1998). In practice, this assumption is rarely valid. Howard (1998) explains this with the following example:

> Do English or Scots cattle have a higher proportion of cows infected with a certain virus? Suppose we were informed (before collecting any data) that the proportion of English cows infected was 0.8. With independent uniform priors we would now give $\mathscr{H}_1$ ($p_1 > p_2$) a probability of 0.8 (…) In very many cases this would not be appropriate. Often we will believe (for example) that if $p1$ is 80%, $p2$ will be near 80% as well and will be almost equally likely to be larger or smaller. (We are still assuming it will never be exactly the same.) (p. 363)

PsychOpen GOLD

The second assumption of the IBE approach is that an effect is always present; that is, training the bees to prefer a certain color may increase approach rates or decrease approach rates; it is never the case that the training is completely ineffective. This assumption follows from the fact that a continuous prior does not assign any probability to a specific point value such as $\delta = 0$ (Jeffreys, 1939; Williams, Bååth, & Philipp, 2017; Wrinch & Jeffreys, 1921). Thus, using the IBE approach practitioners can only test whether the alterations in the experimental group yield a positive or a negative effect. Obtaining evidence in favor of the null hypothesis—which was one of the desiderata listed by Gronau et al. (2021)—is not possible with this approach. Hence, the IBE approach does not represent a testing effort, but rather an estimation effort (Jeffreys, 1939). To allow for both hypothesis testing and parameter estimation a Bayesian A/B testing model has to be able to assign prior mass to the possibility that the difference between the two conditions is exactly zero. It should be acknowledged that this can be achieved when the success probabilities are assigned beta priors (e.g., Günel & Dickey, 1974; Jamil et al., 2017; Jeffreys, 1961); however, here we follow the recent recommendation by Dablander et al. (2022) and adopt an alternative statistical approach.

The Logit Transformation Testing (LTT) Approach

An A/B test model that assigns prior mass to the null hypothesis of no effect was introduced by Kass and Vaidyanathan (1992) and implemented by Gronau et al. (2021). In contrast to the IBE approach, this model assigns a prior distribution to the log odds ratio, thereby accounting for the dependency between the success probabilities of the two experimental groups. The LTT approach is specified as follows:

$$
\begin{aligned}
y_A &\sim \mathit{Binomial}(n_A, \theta_A) \\
y_B &\sim \mathit{Binomial}(n_B, \theta_B) \\
\log\left(\frac{\theta_A}{1-\theta_A}\right) &= \gamma - \psi/2 \\
\log\left(\frac{\theta_B}{1-\theta_B}\right) &= \gamma + \psi/2
\end{aligned}
$$

As before, this model assumes that y_A and y_B follow binomial distributions with success probabilities θ_A and θ_B. However, the success probabilities are a function of two parameters, γ and ψ. Parameter γ indicates the grand mean of the log odds, while ψ denotes the distance between the two conditions (i.e., the log odds ratio; Bland & Altman, 2000; Hailpern & Visintainer, 2003). The hypothesis that there is no difference between the two groups can be formulated as a null hypothesis: $\mathcal{H}_0: \psi = 0$. Under the alternative hypothesis $\mathcal{H}_1$, ψ is assumed to be nonzero. By default, both parameters are assigned normal priors:

PsychOpen GOLD

$$\gamma \sim N\left(\mu_\gamma, \sigma_\gamma^2\right)$$
$$\psi \sim N\left(\mu_\psi, \sigma_\psi^2\right)$$

While the choice of a prior for γ is relatively inconsequential for the comparison between $\mathcal{H}_0$ and $\mathcal{H}_1$, the choice of a prior for ψ is far-reaching: it determines the predictions of $\mathcal{H}_1$ concerning the difference between versions A and B. In other words, ψ is the test-relevant parameter.[6]

We consider four hypotheses that may be of interest in practice:

$\mathcal{H}_0$: $\theta_A = \theta_B$; The success probabilities θ_A and θ_B are identical.

$\mathcal{H}_1$: $\theta_A \neq \theta_B$; The success probabilities θ_A and θ_B are not identical.

$\mathcal{H}_+$: $\theta_B > \theta_A$; The success probability θ_B is larger than the success probability θ_A.

$\mathcal{H}_-$: $\theta_A > \theta_B$; The success probability θ_A is larger than the success probability θ_B.

By comparing these hypotheses, practitioners may obtain answers to the following questions:

1. Is there a difference between the success probabilities, or are they the same? This requires a comparison between $\mathcal{H}_1$ and $\mathcal{H}_0$.
2. Does group B have a higher success probability than group A, or are the probabilities the same? This requires a comparison between $\mathcal{H}_+$ and $\mathcal{H}_0$.
3. Does group A have a higher success probability than group B, or are the probabilities the same? This requires a comparison between $\mathcal{H}_-$ and $\mathcal{H}_0$.
4. Does group B have a higher success probability than group A, or does group A have a higher success probability than group B? This is the question that is also addressed by the IBE approach discussed earlier, and it requires a comparison between $\mathcal{H}_+$ and $\mathcal{H}_-$.

To quantify the evidence that the observed data provide for and against the hypotheses we compare the models' predictive performance.[7] For two models, say $\mathcal{H}_0$ and $\mathcal{H}_+$, the ratio of their average likelihoods for the observed data is known as the *Bayes factor* (Jeffreys, 1939; Kass & Raftery, 1995; Wagenmakers et al., 2018):

$$BF_{+0} = \frac{p\left(data \mid \mathcal{H}_+\right)}{p\left(data \mid \mathcal{H}_0\right)},$$

where BF_{+0} indicates the extent to which $\mathcal{H}_+$ outpredicts $\mathcal{H}_0$.

The evidence from the data is expressed in the Bayes factor, but to compare two hypotheses in their entirety, the *a priori* plausibility of the hypotheses needs to be

6) Note that the overall prior distribution for ψ can be considered a mixture between a 'spike' at 0 coming from $\mathcal{H}_0$ and a Normal 'slab' coming from $\mathcal{H}_1$ (e.g., van den Bergh et al., 2021).

7) We use the terms 'model' and 'hypothesis' interchangeably.

PsychOpen GOLD

considered as well. Bayes' rule describes how we can use the Bayes factor to update the relative plausibility of the two competing models after having seen the data (Kass & Raftery, 1995; Wrinch & Jeffreys, 1921):

$$\underbrace{\frac{p(\mathcal{H}_+ \mid data)}{p(\mathcal{H}_0 \mid data)}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{H}_+)}{p(\mathcal{H}_0)}}_{\text{prior odds}} \times \underbrace{\frac{p(data \mid \mathcal{H}_+)}{p(data \mid \mathcal{H}_0)}}_{BF_{+0}}.$$

The prior odds quantify the plausibility of the hypotheses before seeing the data, while the posterior odds quantify the plausibility of the two hypotheses after taking the data into account (Wagenmakers et al., 2018). The Bayes factor is the evidence—the change from prior to posterior plausibility brought about by the data.

Implementation of the LTT Approach in R and JASP — To demonstrate the analyses with the LTT approach we can use the `abtest` package (version 1.0.1, Gronau, 2019) in R (R Core Team, 2020). The functionality of this package has also been implemented in JASP (JASP Team, 2020). Below we first discuss the R code and then turn to the JASP implementation. Note that the same analysis can also be performed with other more general software for Bayesian inference, such as JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017).

It is recommended that a hypothesis is specified before setting up the A/B test (McFarland, 2012). For the previous example, it can be assumed that the researcher hypothesized that bees may indeed have color vision and that the bees in Group B will therefore approach the blue disc relatively frequently during the testing phase. Hence, from a Bayesian perspective, we may want to compare the directional hypothesis $\mathcal{H}_+$ (i.e., that bees in Group B approach the blue disc more often than bees in Group A) against the null hypothesis $\mathcal{H}_0$ (i.e., there is no difference in the approach rate between Groups A and B). To prevent the undesirable impact of hindsight bias it is likewise recommended to specify the prior distribution for the log odds ratio ψ under $\mathcal{H}_+$ before having inspected the data.

For illustrative purposes we assume that in the present example there is little prior knowledge, which motivates the specification of an uninformed standard normal prior distribution: $\mathcal{H}_1{:}\,\psi \sim N(0, 1)$, which is also the default in the `abtest` package. With the hypotheses of interest specified and the prior distributions assigned to the test-relevant parameter, we are almost ready to execute the Bayesian hypothesis test using the `ab_test` function. This function requires the data, parameter priors, and prior model probabilities. For the present example, we set the prior probabilities of $\mathcal{H}_+$ and $\mathcal{H}_0$ equal to 0.5 and we assign the grand mean parameter γ a relatively uninformative standard normal prior distribution:

PsychOpen GOLD

```
R> library(abtest)
R> bees2 <- as.list(read.csv2("bees_data2.csv")[-1,-1])
R> prior_prob <- c(0, 0.5, 0, 0.5)
R> names(prior_prob) <- c("H1", "H+", "H-", "H0")
R> AB2 <- ab_test(data = bees2, prior_par = list(mu_psi = 0,
+                   sigma_psi = 1, mu_beta = 0, sigma_beta = 1),
+                   prior_prob = prior_prob)
```

As shown in the code above, the standard normal prior on ψ is specified by assigning values for `mu_psi` and `sigma_psi` to the `prior_par` argument of the `ab_test` function. The prior model probabilities are specified by feeding a vector which specifies the probability for the hypotheses $\mathcal{H}_1$, $\mathcal{H}_+$, $\mathcal{H}_-$, and $\mathcal{H}_0$ to the `prior_prob` argument. The `ab_test` function then returns the Bayes factors and the prior and posterior probabilities of the hypotheses. A more detailed explanation of the function and its arguments can be obtained by typing `?ab_test` into the R console. For the bee example, the Bayes factor BF_{+0} equals 4.7, meaning that the data are approximately 5 times more likely under the alternative hypothesis $\mathcal{H}_+$ than under the null hypothesis $\mathcal{H}_0$. A Bayes factor of ~ 5 is generally considered moderate evidence (e.g., Jeffreys, 1939; Lee & Wagenmakers, 2013).

The robustness of this conclusion can be explored by changing the prior distribution on ψ (i.e., by varying the mean and standard deviation of the normal prior distribution) and observing the effect on the Bayes factor. Figure 5 visualizes the robustness of the Bayes factor for changes across a range of values for μ_ψ and σ_ψ. The Bayes factor is highest for low σ_ψ values and $\mu_\psi \approx 0.6$. The heatmap shows that our conclusion regarding the evidence for $\mathcal{H}_+$ over $\mathcal{H}_0$ is relatively robust. The plot can be produced with:

```
R> plot_robustness(AB2, mu_range = c(0, 2), sigma_range = c(0.1, 1),
+                   bftype = "BF+0")
```

A sequential analysis tracks the evidence in chronological order. Figure 6 shows how the posterior probability of either hypothesis unfolds as the observations accumulate. The figure indicates that after some initial fluctuations, and a tie after about 90 observations, the last 110 observations cause the probability for the alternative hypothesis to increase steadily until it reaches its final value of 0.826. Because we consider only two hypotheses, the probability for $\mathcal{H}_+$ is the complement of that for $\mathcal{H}_0$. The sequential analysis can be obtained as follows:

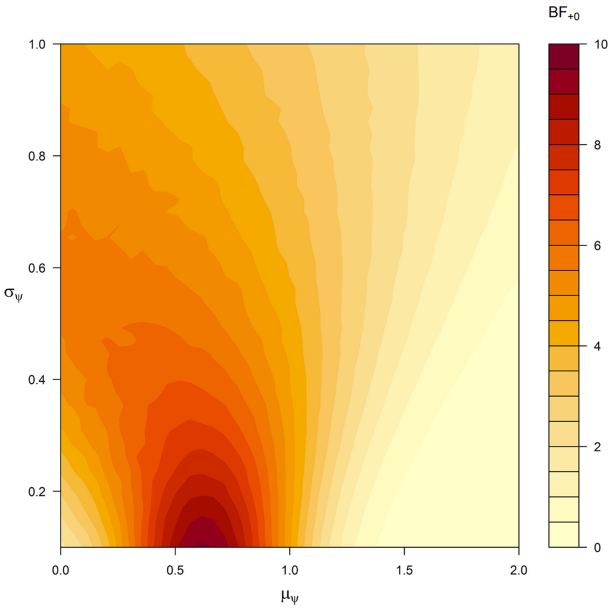
```
R> plot_sequential(AB2)
```

Figure 6 also visualizes the prior and posterior probabilities of the hypotheses as a probability wheel. The probability of $\mathcal{H}_+$ has increased from 0.5 to 0.826 while the posterior plausibility of $\mathcal{H}_0$ has correspondingly decreased from 0.5 to 0.174.

PsychOpen GOLD

Figure 5

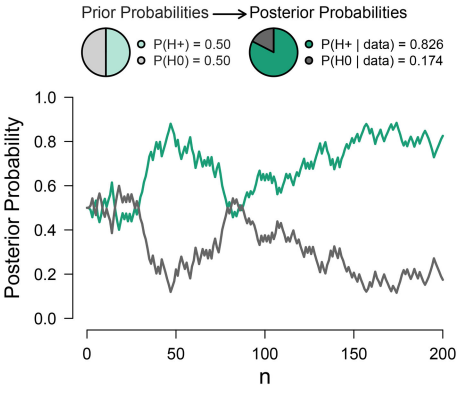
Bayes Factor Robustness Plot for the Comparison Between H_0 and H_+ as Applied to the Fictitious Bee Data (i.e., A = 50/100 Versus B = 65/100)



Note. The highest Bayes factor is reached for a prior on ψ with $\mu_\psi \approx 0.6$. The conclusion that there is moderate evidence for H_+ over H_0 holds across a large range of values for μ_ψ and σ_ψ.

Figure 6

The Flow of Posterior Probability for H_0 and H_+ as a Function of the Accumulating Number of Observations for the Fictitious Bee Data (i.e., A = 50/100 vs. B = 65/100)



PsychOpen GOLD

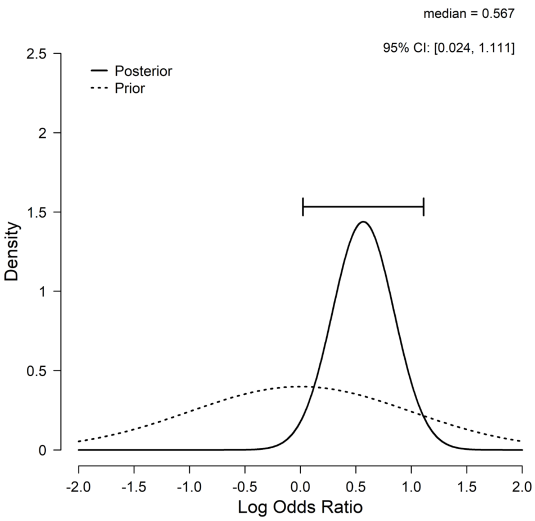
Having collected evidence for the hypothesis that trained bees prefer the blue disc more than do untrained bees, one might then wish to quantify the size of this difference in preference. To do so, we switch from a testing framework to an estimation framework. For this purpose, we adopt the two-sided model $\mathscr{H}_1$ and use Bayes' rule to obtain the posterior distribution for the log odds ratio. Figure 7 shows the result as produced via the `plot_posterior` function:

```
R> plot_posterior(AB2, what = 'logor')
```

The dotted line in Figure 7 displays the prior distribution, the solid line displays the posterior distribution (with a 95% central credible interval [CI]), and the posterior median and 95% CI are displayed on top. For our fictitious bee example, Figure 7 indicates that the log odds ratio is 95% probable to lie between 0.024 and 1.111. It is important to realize that this inference is conditional on $\mathscr{H}_1$ (van den Bergh et al., 2021), which features a prior distribution that makes two strong assumptions: (1) the effect is either positive or negative, but never zero; (2) a priori, effects are just as likely to be positive as they are to be negative (i.e., the prior distribution for the log odds ratio is symmetric around zero).

Figure 7

Prior and Posterior Distribution of the Log Odds Ratio Under H_1 as Applied to the Fictitious Bee Data (i.e., A = 50/100 vs. B = 65/100)



Note. The posterior median and the 95% credible interval are shown in the top right corner.

The posterior distributions for the two success probabilities can be inspected separately using:

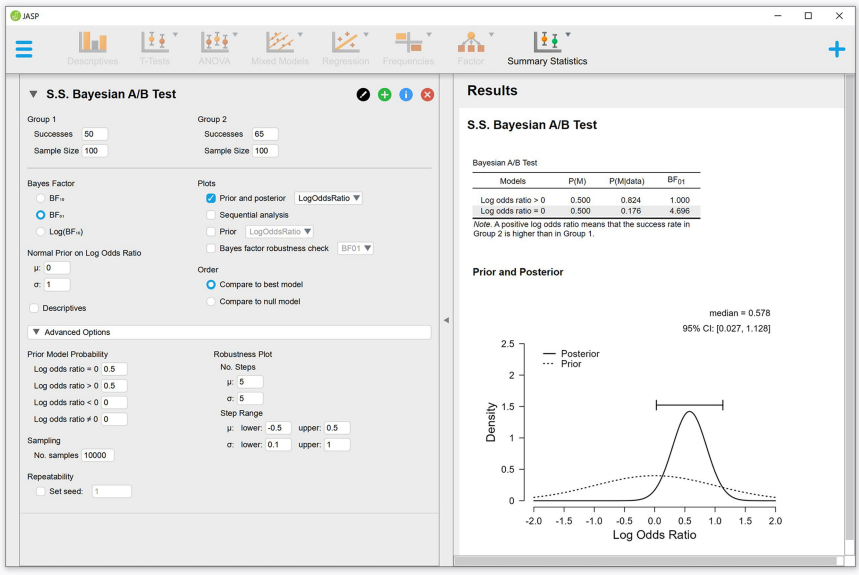
PsychOpen GOLD

```
R> plot_posterior(AB2, what = 'p1p2')
```

The `abtest` R package has also been implemented in JASP, allowing teachers, students, and researchers to obtain the above results with a graphical user interface. A screenshot is provided in Figure 8. There are two ways in which the `abtest` functionality can be activated in JASP. The first method, shown in Figure 8, is to activate the *Summary Statistics* module using the blue '+' sign in the top right corner. Clicking the *Summary Statistics* icon on the ribbon and selecting `Frequencies → Bayesian A/B Test` brings up the interface shown in Figure 8.

Figure 8

JASP (Version 0.16.3.0) Screenshot of the Summary Statistics Implementation of the `abtest` R Package



Note. This features the comparison between H_0 and H_+ as applied to the fictitious bee data (i.e., A = 50/100 versus B = 65/100). The left panel shows the input options and the right panel shows the associated analysis output.

Using the *Summary Statistics* module, users only need to enter the total number of successes and sample sizes in the two groups. As shown in Figure 8, the input panel offers the similar functionality to the `abtest` R package. The slight difference in outcomes is due to the fact that the results for the directional hypotheses $\mathscr{H}_+$ and $\mathscr{H}_-$ involve importance sampling (Gronau et al., 2021).

The second method to activate the `abtest` functionality in JASP is to store the results in a data file, open it in JASP, click the *Frequencies* icon on the ribbon and then

select `Bayesian → A/B Test`. When the data file contains the intermediate results, this second method allows users to conduct a sequential analysis such as the one shown in Figure 6.

To showcase the different approaches to Bayesian A/B testing we now apply the methodology to two example data sets.[8] The first example data set features real data collected on the 'Rekentuin' online learning platform, and the second example data set features fictitious data constructed to be representative of online webshop experiments (i.e., relatively small effect sizes and relatively high sample sizes).

Example I: The Rekentuin

The Rekentuin A/B Experiment

Rekentuin (Dutch for 'math garden') is a tutoring website where children can practice their arithmetic skills by playing adaptive online games. The Rekentuin website is visited by Dutch elementary school children between the ages of 4 and 12. During the testing interval from the 22nd of January 2019 to the 5th of February 2019, a total of 15,322 children were active on Rekentuin.

Figure 9

Screenshots from the Rekentuin Web Environment

Note. The left-hand panel shows a screenshot of a Rekentuin landing page. The page shows that the child has earned three crowns for the category 'optellen' (Dutch for 'addition'). The right-hand panel shows a screenshot of an addition problem in the Rekentuin web environment. The coins at stake are displayed on the bottom right corner.

The left-hand panel of Figure 9 shows a screenshot of a Rekentuin landing page. In Rekentuin, children earn coins by quickly solving simple arithmetic problems that are

8) For general recommendations on how to apply Bayesian procedures and interpret the results, consider van Doorn et al. (2021).

PsychOpen GOLD

organized into different classes (e.g., addition, subtraction, division, etc.). An example of an addition problem is shown in the right-hand panel of Figure 9, with the coins at stake shown in the bottom right corner. The children can use the coins that are gained to buy virtual trophies (not shown). The better a given child performs, the more trophies they are able to add to their trophy cabinet. The prospect of earning trophies motivates the children to participate and perform well (for details see Brinkhuis et al., 2018; Klinkenberg, Straatemeier, & van der Maas, 2011). On the Rekentuin landing page, the plant growth near each class of arithmetic problem indicates the extent to which that class was recently practiced; practice makes the plants grow, whereas periods of inactivity makes the plants wither away.

In 2019, the developers of Rekentuin faced the challenge that many children would preferentially engage with the class of arithmetic problems that they had already mastered (e.g., addition)—a sensible strategy if the goal is to maximize the number of coins gained. To incentivize the children to practice other classes of arithmetic problems (e.g., subtraction) the developers implemented a 'crown' for the type of games that the children had already mastered (see Figure 9, left-hand panel). Children could gain up to three crowns for each type of game. Thus, in order to obtain more crowns, children had to engage more frequently with the types of games they had played less often. However, the crowns did not have the desired effect—instead of decreasing the playtime on crown games, the playtime on crown games actually increased.

To induce the children to play other games, the Rekentuin developers constructed a less subtle manipulation: they removed the virtual reward (i.e., the coins) from the crown games. To test the effectiveness of this manipulation, the Rekentuin developers designed an A/B test. Half of the children continued playing on an unchanged website (Version A), whereas the other half could no longer earn coins for crown games (Version B). The children playing Version B were not notified of the change but had to discover the changes for themselves.

The question of interest is whether changing the incentive structure for crown games (i.e., removing the coins) had the desired effect. To address this question we analyzed the Rekentuin data set using the two Bayesian A/B testing approaches outlined earlier.

Method

Preregistration

The data were collected by Abe Hofman and colleagues on the Rekentuin website in 2019. All intended analyses were applied to synthetic data and the associated analysis scripts were stored on a repository at the OSF. We did not inspect the data before the preregistration was finalized. All preregistration materials as well as the real data are available in the Supplementary Materials section.

PsychOpen GOLD

Data Preprocessing

Our analysis concerns the last game that each child played during the testing interval: was it a crown game or not? By examining only the last game we obtain a binary variable (required for the present A/B test) and also allow children the maximum opportunity to experience that crown games no longer yield coins.

We excluded children from the analyses according to two criteria. Firstly, we excluded 8573 children who did not play any crown game during the time of testing because they could not have experienced the experimental manipulation in Version B. Secondly, we excluded 350 children who only played one crown game and it was their last game, because for these children we cannot observe the potential influence of the manipulation on their playing behavior. In total, we therefore excluded 8923 children.

Results

Descriptives

The Rekentuin data are summarized in Table 1. The table indicates the number of children who played a crown game or a non-crown game as their last game. In the control condition, 2272 out of 3178 children (≈ 71.5%) played a non-crown game as their last game; in the treatment condition, with the coins for crown games removed, this was the case for 2596 out of 3221 children (≈ 80.6%). It appears the manipulation had a large effect. We now quantify the statistical evidence using the Bayesian A/B test.

Table 1

Descriptives of the Rekentuin Data

	Game Type		
Coins	**Non-Crown**	**Crown**	**Total**
Yes	2272	906	3178
No	2596	625	3221
Total	4868	1531	6399

Note. Children in Version B (no coins available in crown games) played more non-crown games.

Rekentuin A/B Test: The IBE Approach

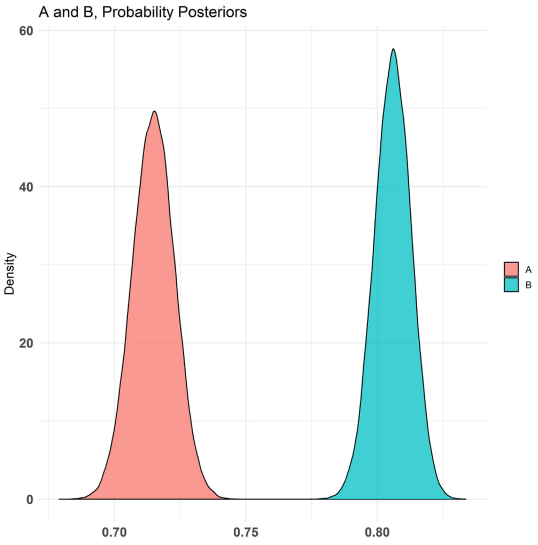
As before, in the IBE approach we assigned two uninformed beta(1,1) distributions to the success probabilities of Versions A and B.[9] Figure 10 displays the resulting two independent posterior distributions.

9) Researchers with access to pre-intervention data could instead consider to use an informed prior distribution, although there is always a risk that the pre-intervention data differ from the post-intervention data on some unknown dimension.

PsychOpen GOLD

Figure 10

Independent Posterior Beta Distributions of the Success Probabilities of Playing a Non-Crown Game



Note. Version A corresponds to the unchanged Rekentuin website. Version B denotes the Rekentuin version where the children could not earn coins for crown games. The plot is produced by the `bayesAB` package. The analysis used two independent beta(1, 1) priors.

Consistent with the intuitive impression from Table 1, virtually all of the posterior mass in Version B is on higher values of $\theta^{\text{non-crown}}$ than that in Version A. This suggests that the success probability of the modified Rekentuin version is higher and that removing the coins from the crown games had a positive impact on the number of non-crown games played.

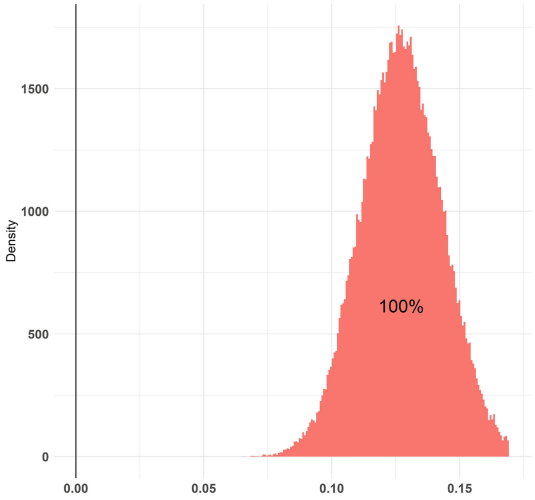
Figure 11 shows the conversion rate uplift. The distribution peaks at around 0.12, indicating that the most likely conversion increase equals 12%. Also, all posterior mass (i.e., 100% of the samples) is above zero. In other words, we can be relatively certain that Version B is better than Version A. Note that this statement assumes that it is equally likely that the alterations in Version B had a positive or negative effect on the rate at which the children played non-crown games.

In addition to the `bayesAB` package output, we computed the posterior probability of the event $\theta_B^{non-crown} > \theta_A^{non-crown}$ using the formula reported by Schmidt and Mørup (2019). For the Rekentuin data, $p(\theta_B > \theta_A) \approx 1$. The analytic calculation of the posterior distribution for the difference $\delta = \theta_B^{non-crown} - \theta_A^{non-crown}$ between the two independent beta distributions fails because the data set is too large. Figure 12 plots the entire probability distribution of the difference δ calculated using the normal approximation. The distribution is very narrow and peaks at around 0.09.

PsychOpen GOLD

Figure 11

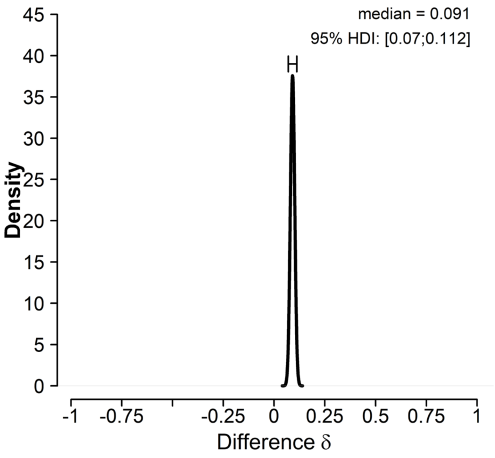
Histogram of the Conversion Rate Uplift from Version A (i.e., 2272/3178) to Version B (i.e., 2596/3221) for the Rekentuin Data



Note. The conversion rate indicates the proportion of children that played a non-crown game. The uplift is calculated by dividing the difference in conversion by the conversion in A. The plot is produced by the `bayesAB` package.

Figure 12

Posterior Distribution of the Difference $\delta = \theta_B^{\text{Non-Crown}} - \theta_A^{\text{Non-Crown}}$ *for the Proportion of Non-Crown Games Between the Two Rekentuin Website Versions*

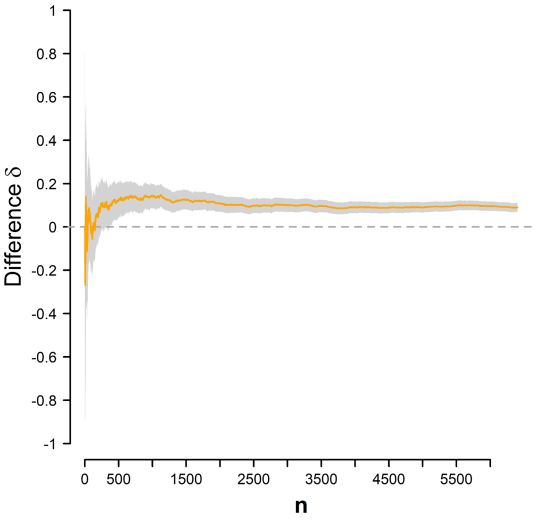


Note. Children in Version B—the modified website version—played more non-crown games compared to children playing on the website Version A.

PsychOpen GOLD

Figure 13

Sequential Analysis of the Difference Between the Success Probabilities (i.e., $\theta_B^{Non-Crown} - \theta_A^{Non-Crown}$) of the two Rekentuin Versions



Note. The orange line plots the posterior mean of the difference. The grey area visualizes the width of the highest density interval as a function of sample size *n*.

Figure 13 plots the sequential analysis of the posterior mean of the difference between $\theta_A^{non-crown}$ and $\theta_B^{non-crown}$ as well as the 95% HDI of the difference. After some initial fluctuation, the posterior mean difference between $\theta_A^{non-crown}$ and $\theta_B^{non-crown}$ settles at ~0.09 while the HDI becomes more narrow with increasing sample size. The range of likely values for δ eventually ranges from approximately 0.071 to 0.112.

Rekentuin A/B Test: The LTT Approach

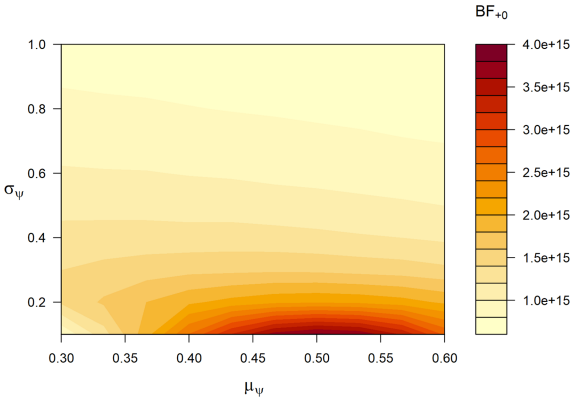
For the LTT approach, we compare the directional hypothesis $\mathcal{H}_+$ (i.e., children in Version B play more non-crown games then children in Version A) against the null hypothesis $\mathcal{H}_0$ (i.e., the proportion of non-crown games played does not differ between Versions A and B). We employed a truncated normal distribution with $\mu = 0$ and $\sigma^2 = 1$ under the alternative hypothesis as there is a range of parameter values that seem plausible (see, for example, Cameron, Banko, & Pierce, 2001; Tang & Hall, 1995). In particular, it is plausible that removing the coins from the crown games results in a marked change.

The observed sample proportions of 0.806 for Version B and 0.715 for Version A suggest that the children in Version B played more non-crown games as compared to Version A. The Bayes factor BF_{+0} that assesses the evidence in favor of our hypothesis

PsychOpen GOLD

Figure 14

Bayes Factor Robustness Plot for the Rekentuin Data



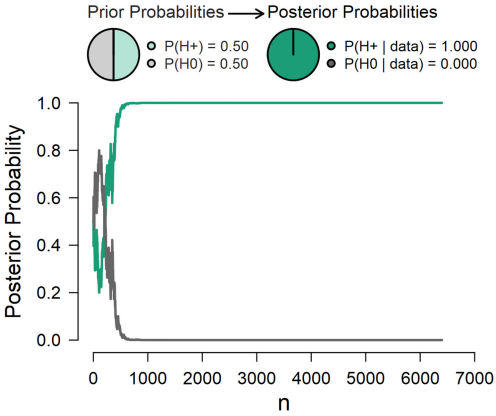
that the children in Version B played more non-crown games equals 7.944e+14. This means that the data are about 800 trillion times more likely to occur under the alternative hypothesis $\mathscr{H}_+$ than under the null hypothesis $\mathscr{H}_0$. In sum, the Bayes factor indicates overwhelming evidence for the alternative hypothesis (e.g., Jeffreys, 1939; Lee & Wagenmakers, 2013). Figure 14 visualizes the dependency of the Bayes factor on the prior distribution for ψ by varying the mean μ_ψ and standard deviation σ_ψ of the normal prior distribution. From looking at the heatmap we can conclude that the Bayes factor is robust. The data indicate extreme evidence across a range of different values for the prior distribution on ψ.

Figure 15 tracks the evidence for either hypothesis in chronological order. After about 800 observations, the evidence for $\mathscr{H}_+$ is overwhelming. The posterior probabilities of the hypotheses are also shown as a probability wheel on the top of Figure 15. The green area visualizes the posterior probability of the alternative hypothesis and the grey area visualizes the posterior probability of the null hypothesis. The data have increased the plausibility of $\mathscr{H}_+$ from 0.5 to almost 1 while the posterior plausibility of the null hypothesis $\mathscr{H}_0$ has correspondingly decreased from 0.5 to almost 0.

PsychOpen GOLD

Figure 15

The Flow of Posterior Probability for H_0 and H_+ as a Function of the Number of Observations Across Both Rekentuin Versions



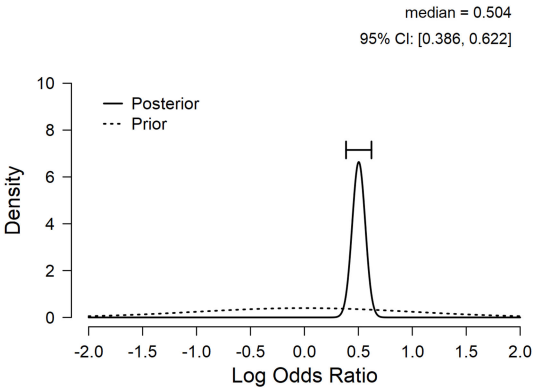
Note. The prior and posterior probabilities of the hypotheses are displayed on top.

In sum, the evidence in favor of the alternative hypothesis is overwhelming. To complete the picture, we quantified the difference between the two Rekentuin versions by estimating the size of the log odds ratio. Figure 16 shows the prior and posterior distribution for the log odds ratio under the two-sided model $\mathscr{H}_1$. The dotted line displays the prior distribution and the solid line displays the posterior distribution (with 95% central CI). The plot indicates that, given that the log odds ratio is not exactly zero, it is 95% probable to lie between 0.386 and 0.622, where the posterior median is 0.504. In R and in JASP, this prior and posterior plot may also be shown on a different scale: as an odds ratio, relative risk, absolute risk, and individual proportions.

PsychOpen GOLD

Figure 16

Prior and Posterior Distribution of the Log Odds Ratio Under H_1 for the Rekentuin Data Set



Note. The median and the 95% credible interval of the posterior density for the Rekentuin data are shown in the top right corner.

Example II: The Fictional Webshop

The Rekentuin manipulation directly targeted children's motivation to play the games. Common A/B tests for web development purposes implement more subtle manipulations that result in much smaller effect sizes. In this section we analyze such a scenario.

Consider the following fictitious scenario: an online marketing team seeks to improve the click rate on a call-to-action button on their website's landing page. Therefore, they devise an A/B test. Half of the website visitors read 'Try our new product!' (Version A), and the other half reads 'Test our new product!' (Version B).[10] The success of the website versions is measured by the rate at which website visitors click on the call-to-action button.

To demonstrate the analyses we use synthetic data. The corresponding R code can be found at the Supplementary Materials. Table 2 provides the number of clicks in each group. The conversion rate equals 1131/10000 = 0.1131 in Version A and 1275/10000 = 0.1275 in Version B. The company now wishes to determine whether and to what extent the observed sample difference in proportions translates to the population.

10) This example was inspired by a real conversion rate optimization project at https://blog.optimizely.com/2011/06/08/optimizely-increases-homepage-conversion-rate-by-29/

PsychOpen GOLD

Table 2

Descriptives of the Fictional Webshop Data

	Click on Button		
Condition	**Yes**	**No**	**Total**
A	1131	8869	10000
B	1275	8725	10000
Total	2406	17594	20000

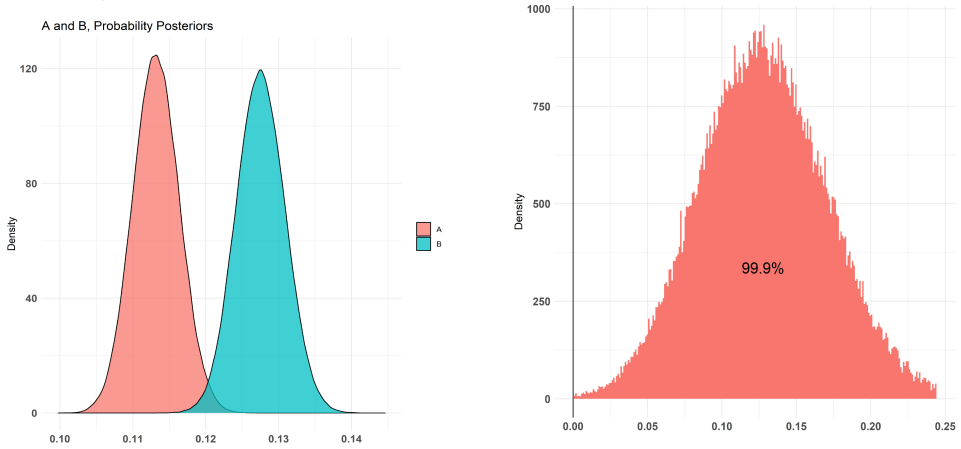
Note. Visitors confronted with Version B clicked the call-to-action button more often than those confronted with Version A.

The IBE Approach

We again use the `bayesAB` package in R to analyze the data according to the IBE approach using the default independent beta(1, 1) distributions on θ_A and θ_B (Portman, 2017; R Core Team, 2020).

Figure 17

Two Independent Posterior Distributions and Conversion Rate Uplift



Note. The left-hand panel shows the independent posterior beta distributions of the click-through probabilities for fictional webshop Versions A and B. The plot is produced by the `bayesAB` package. The right-hand panel shows histogram of the conversion rate uplift from Version A to Version B for the fictional webshop data.

The left-hand panel of Figure 17 illustrates the two independent posterior distributions. We can see that Version B's posterior distribution for the success probability assigns more mass to higher values of θ. This suggests that the click-through rate for Version

PsychOpen GOLD

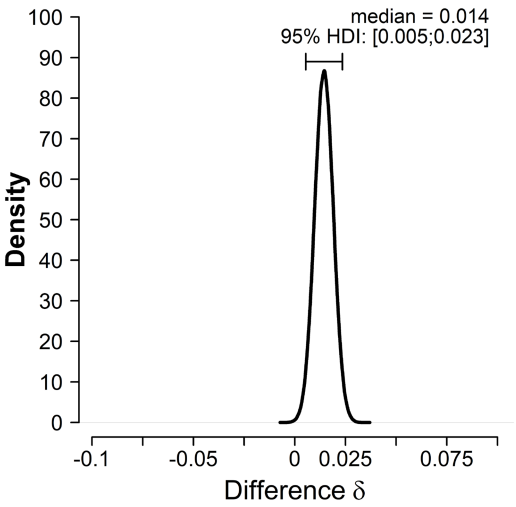
B's message 'Test our new product!' is higher than that for Version A's message 'Try our new product!'.

The right-hand panel of Figure 17 depicts the conversion rate uplift. The posterior distribution for the uplift peaks at around 0.125, indicating that the most likely conversion increase equals 12.5%. Also, most posterior mass (i.e., 99.9% of the samples) is above zero, indicating that we can be 99.9% certain that Version B is better than Version A instead of the other way around.

The analytically calculated posterior probability of the event $\theta_B > \theta_A$ equals 0.999 (Schmidt & Mørup, 2019). Figure 18 shows the posterior difference distribution δ. The distribution peaks at 0.014. A difference of this size is relatively large for a conversion rate optimization endeavor (Browne & Jones, 2017). We calculated δ for this data set with the normal approximation.

Figure 18

Posterior Distribution of the Difference $\delta = \theta_B - \theta_A$ for the Click-Through Proportion Between the two Fictitious Website Versions



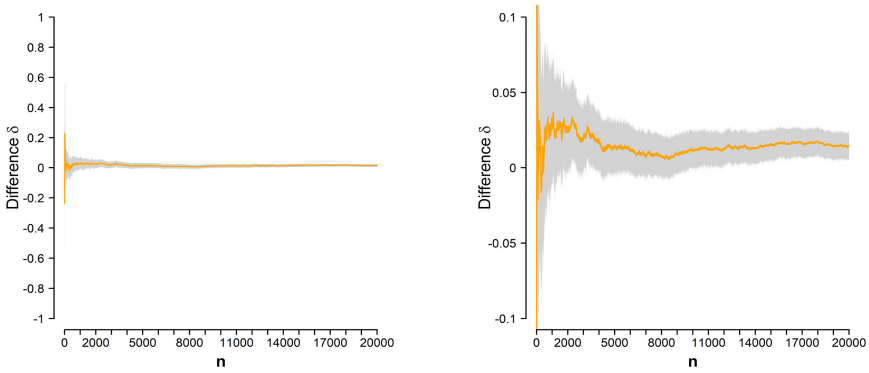
Note. Visitors confronted with Version B clicked more often on the call-to-action button than visitors confronted with Version A.

Figure 19 plots the posterior mean of the difference between θ_A and θ_B as well as the 95% HDI of the difference in a sequential manner. With increasing sample size, the HDI becomes more narrow. This indicates that the range of likely values for δ becomes smaller. After some initial fluctuation, the posterior mean difference between the two success probabilities θ_A and θ_B settles at ~ 0.014.

PsychOpen GOLD

Figure 19

Sequential Analysis of the Difference Between the Click-Through Probabilities (i.e., $\theta_B - \theta_A$) of the two Fictitious Webshop Versions



Note. The orange line plots the posterior mean of the difference. The grey area visualizes the width of the highest density interval as a function of sample size *n*. The left-hand panel shows the sequential analysis of the difference, with the y-axis ranging from -1 to 1. The right-hand panel shows the sequential analysis of the difference, with the y-axis ranging from -0.1 to 0.1.

The LTT Approach

Before the data can be analyzed according to the LTT approach, a prior distribution for the log odds ratio has to be specified. For this purpose, it is important to note that the subtle manipulations of common A/B tests generally result in very small effect sizes. The effect size of website changes (i.e., the difference in conversion rates between the baseline version and its modification) is typically as small as 0.5% or less (Berman et al., 2018). This means that the analysis of such data requires an exceptionally narrow prior distribution that peaks at a value close to 0 in order for the shape of the prior distribution to do justice to the relative plausibility of parameter values.

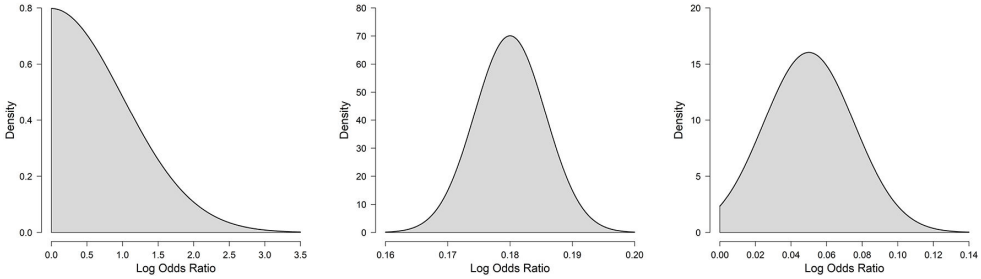
For the present example, we will compare the impact of different prior distributions on the analysis outcome (i.e., a sensitivity analysis). Suppose that the online marketing team specifies three prior distributions. Firstly there is the prior distribution specified by a team member who is still relatively unfamiliar with conversion rate optimization. This team member lacks substantive knowledge about plausible values of the log odds ratio and prefers to use the uninformed standard normal prior, truncated at zero to represent the expectation of a positive effect (see Figure 20, left-hand panel). The two remaining prior distributions come from team members with prior knowledge on A/B testing; consequently, these distributions are much more narrow. One prior comes from a team member who is optimistic about the conversion rate increase in Version B. Specifically,

PsychOpen GOLD

this team member believes the most likely value of the odds ratio to be around 1.20 (i.e., $\mu_\psi = 0.18$; see Figure 20, center panel). The final prior comes from a team member who believes the most likely value of the odds ratio to be around 1.05 (i.e., $\mu_\psi = 0.05$; see Figure 20, right-hand panel). All three prior distributions are truncated at zero to specify a positive effect. In sum, the data will be analyzed with three priors: the uninformed prior with $\mu_\psi = 0$ and $\sigma_\psi = 1$, the optimistic prior with $\mu_\psi = 0.18$ and $\sigma_\psi = 0.005$, and the conservative prior with $\mu_\psi = 0.05$ and $\sigma_\psi = 0.03$.

Figure 20

Three Different Prior Distributions for the Analysis of the Fictitious Webshop Data



Note. The uninformed prior (left-hand panel), $\mathcal{H}_+^u : \psi \sim N^+(0, 1)$. The optimistic prior (center panel), $\mathcal{H}_+^o : \psi \sim N^+(0.18, 0.005^2)$. The conservative prior (right-hand panel), $\mathcal{H}_+^c : \psi \sim N^+(0.05, 0.03^2)$.

We used the `abtest` package (Gronau, 2019) in R (R Core Team, 2020) and JASP for the present analysis. Before estimating the size of the effect, we first evaluate the evidence that there is indeed a difference between Version A and B. Overall the data support the hypothesis that visitors confronted with Version B click on the call-to-action button more often than those confronted with Version A. Specifically, compared to the null hypothesis $\mathcal{H}_0$, the data are about 11 times more likely under the uninformed alternative hypothesis $\mathcal{H}_+^u$, about 80 times more likely under the optimistic alternative hypothesis $\mathcal{H}_+^o$, and about 27 times more likely under the conservative alternative hypothesis $\mathcal{H}_+^c$.[11]

The influence of the prior distribution on the Bayes factor can be explored more systematically with the Bayes factor robustness plot, shown in Figure 21. Varying both the mean μ_ψ and the standard deviation σ_ψ of the prior distribution on ψ shows that the BF_{+0} mostly ranges from about 10 to about 60. The evidence is generally less compelling for prior distributions that are relatively wide (i.e., high σ_ψ) or relatively peaked but away from zero (i.e, low σ_ψ and high μ_ψ). In both scenarios, substantial predictive mass

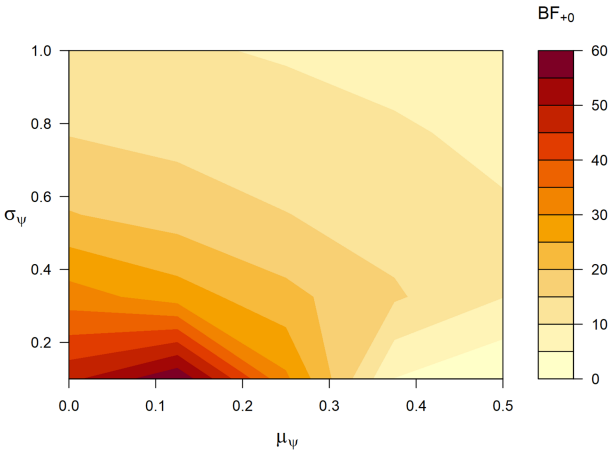
11) It follows from transitivity that the optimistic colleague outpredicted the pessimistic colleague by a factor of 80/27 ≈ 2.96.

PsychOpen GOLD

is wasted on effect sizes that are unreasonably large, and were unlikely to manifest themselves in the context of the present webshop A/B experiment.

Figure 21

Bayes Factor Robustness Plot for the Fictitious Webshop Data



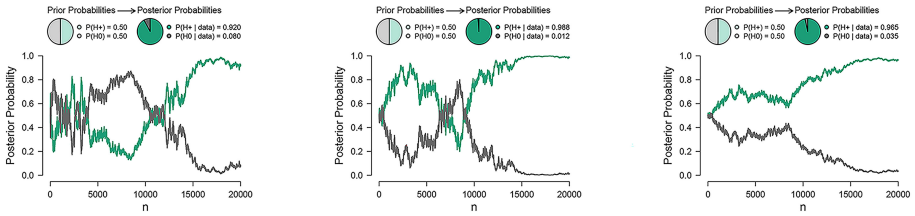
Note. Overall, there is strong evidence for H_1 over H_0 across a range of reasonable values for μ_ψ and σ_ψ. The evidence is less compelling when the prior for the log odds ratio is relatively wide (i.e., when σ_ψ is relatively high) or far away from zero (i.e., when μ_ψ is relatively high).

The prior and posterior probabilities of the hypotheses are displayed on top of Figure 22. For the uninformed prior, the optimistic prior, and the conservative prior, the posterior probabilities for $\mathscr{H}_+$ are approximately equal to 0.920, 0.988, and 0.965, respectively. This illustrates that even though the three priors provide different levels of evidence as measured by the Bayes factor, the overall interpretation is approximately the same. Figure 22 also shows the flow of posterior probability for each of the three prior distributions as a function of the fictitious incoming observations. For all distributions, a clear and consistent pattern of preference starts to emerge only after 10,000 observations, which is when the posterior probability of $\mathscr{H}_+$ gradually rises while that of $\mathscr{H}_0$ decreases accordingly.

PsychOpen GOLD

Figure 22

Flow of Posterior Probability for H_0 and H_+ as a Function of the Number of Observations Across Both Fictitious Website Versions.



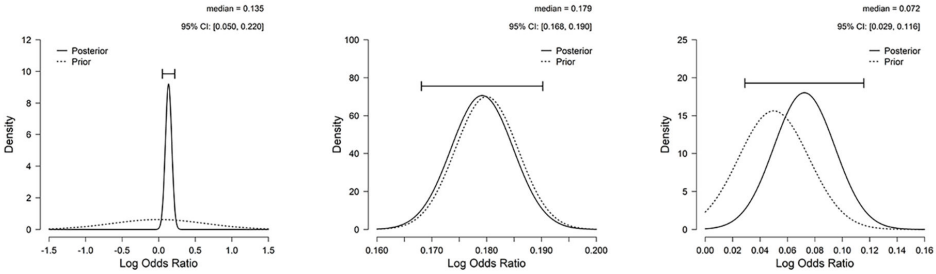
Note. The left-hand panel shows the sequential analysis with the uninformed prior. The center panel shows the sequential analysis with the optimistic prior. The right-hand panel shows the sequential analysis with the conservative prior.

In sum, the fictional webshop data present strong to very strong evidence for the claim that the conversion rate is higher in Version B than in Version A (Lee & Wagenmakers, 2013). Being assured that the effect is present, the online marketing team now wishes to assess the size of the effect. Figure 23 displays the prior and posterior distribution for the log odds ratio using the three different priors. The prior distribution is plotted as a dotted line and the posterior distribution as a solid line (with 95% central CI). Under the assumption that the effect is non-zero, the left-hand panel of Figure 23 indicates that the posterior median of the log odds ratio is 0.135 with a 95% CI ranging from 0.050 to 0.220 when using the uninformed prior distribution. The middle panel of Figure 23 displays the posterior distribution for the log odds ratio when using the optimistic prior distribution. The leftward shift of the posterior distribution indicates that the effect is somewhat smaller than expected; the posterior median is 0.179 and the 95% CI ranges from 0.168 to 0.190. The change from prior to posterior distribution is only modest, and this reflects the fact that the optimistic prior was also relatively peaked, meaning that the prior belief in the relative plausibility of the different parameter values was very strong. Finally, the right-hand panel of Figure 23 displays the posterior distribution when using the conservative prior distribution. The rightward shift of the posterior distribution indicates that the effect is somewhat larger than expected; the posterior median is 0.072 and the 95% CI ranges from 0.029 to 0.116. The general pattern in Figure 23 is that the change from prior to posterior is more pronounced when prior knowledge is weak.

PsychOpen GOLD

Figure 23

Prior and Posterior Distribution of the Log Odds Ratio Under H1 for the Fictitious Webshop Data Set



Note. The median and the 95% credible interval of the posterior density for the fictitious webshop data are shown in the top right corner. The right-hand panel shows the uninformed prior and the posterior distribution of the log odds ratio under $\mathscr{H}_1$. The center panel shows the optimistic prior and the posterior distribution of the log odds ratio under $\mathscr{H}_1$. The right-hand panel shows the conservative prior and the posterior distribution of the log odds ration under $\mathscr{H}_1$.

Concluding Comments

The A/B test concerns a comparison between two proportions and it is ubiquitous in medicine, psychology, biology, and online marketing. Here we outlined two Bayesian A/B tests: the 'Independent Beta Estimation' or IBE approach that assigns independent beta priors to the two proportion parameters, and the 'Logit Transformation Testing' or LTT approach that assigns a normal prior to the log odds ratio parameter. These approaches are based on different assumptions and hence ask different questions. We believe that the LTT approach deserves more attention: in many situations, the assumption of independence for the proportion parameters is not realistic. Moreover, only with the LTT approach is it possible for practitioners to obtain evidence in favor of or against the null hypothesis.[12] Both approaches allow practitioners to monitor the evidence as the data accumulate, and to take prior/expert knowledge into account.

The LTT approach could be extended to include the possibility of an interval-null or perinull hypothesis to replace the traditional point-null hypothesis (Ly & Wagenmakers, 2021; Morey & Rouder, 2011). If the interval is wide, and if $\mathscr{H}_1$ is defined to be non-overlapping (such that the parameter values inside the null-interval are excluded from $\mathscr{H}_1$) then the evidence in favor of the interval-null hypothesis may increase at a much faster rate than that in favor of the point-null (see also Jeffreys, 1939, pp. 196, 197; Johnson & Rossell, 2010). Interval-null hypotheses are particularly attractive in fields

12) It is possible to expand the IBE approach and add a null hypothesis that both success probabilities are exactly equal (e.g., Jeffreys, 1961), yielding an Independent Beta Testing (IBT) approach. A discussion of the IBT is beyond the scope of this paper (cf. Dablander et al., 2022).

PsychOpen GOLD

such as medicine and online marketing, where the purpose of the experiment concerns a practical question regarding the effectiveness of a particular treatment or intervention. An effect size that is so small as to be practically irrelevant will, with a large enough sample, still give rise to a compelling Bayes factor against the point-null hypothesis. This concern can to some extent be mitigated by considering not only the Bayes factor, but also the posterior distribution. In the above scenario, the conclusion would be that an effect is present, but that it is very small.

Despite its theoretical advantages, the Bayesian LTT approach has been applied to empirical data only sporadically. This issue is arguably due to the fact that many researchers are not familiar with this procedure and the practical advantages that it entails. The fact that the LTT approach had, until recently, not been implemented in easy-to-use software is another plausible reason for its widespread neglect. In this manuscript we outlined the Bayesian LTT approach and showed how implementations in R and JASP make it easy to execute. In addition, we demonstrated with several examples how the LTT approach yields informative inferences that may usefully supplement or supplant those from a traditional analysis.

Funding: This research was supported by the Netherlands Organisation for Scientific Research (NWO; grant #016.Vici.170.083).

Acknowledgments: The authors are grateful to Oefenweb for allowing them to analyze the anonymized data and make it publicly available. The use and publication of the anonymized Rekentuin data has been coordinated with and permitted by Oefenweb.

Competing Interests: E. J. W. declares that he coordinates the development of the open-source software package JASP (https://jasp-stats.org), a non-commercial, publicly-funded effort to make Bayesian and non-Bayesian statistics accessible to a broader group of researchers and students.

Data Availability: Data is freely available at Supplementary Materials.

Supplementary Materials

The supplementary materials provided are the data, all preregistration materials, and an online appendix and can be accessed in the Index of Supplementary Materials below.

Index of Supplementary Materials

Hoffmann, T., Hofman, A., & Wagenmakers, E. J. (2022). *Supplementary materials to "Bayesian tests of two proportions: A tutorial with R and JASP"* [Data, preregistration materials, online appendix]. OSF. https://osf.io/anvg2.

PsychOpen GOLD

References

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ, 311*, Article e485. https://doi.org/10.1136/bmj.311.7003.485

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Institute of Mathematical Statistics.

Berman, R., Pekelis, L., Scott, A., & Van den Bulte, C. (2018). *P-hacking and false discovery in A/B testing*. SSRN. https://doi.org/10.2139/ssrn.3204791

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *BMJ, 320*, Article e1468. https://doi.org/10.1136/bmj.320.7247.1468

Brinkhuis, M. J., Savi, A. O., Hofman, A. D., Coomans, F., van Der Maas, H. L., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics, 5*(2), 29–46. https://doi.org/10.18608/jla.2018.52.3

Browne, W., & Jones, M. S. (2017). *What works in e-commerce: A meta-analysis of 6700 online experiments*. Qubit Digital. https://bit.ly/3zy5ycU

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *Behavior Analyst, 24*(1), 1–44. https://doi.org/10.1007/BF03392017

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Dablander, F., Huth, K., Gronau, Q. F., Etz, A., & Wagenmakers, E.-J. (2022). A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions. *Statistics in Medicine, 41*(8), 1319–1333. https://doi.org/10.1002/sim.9278

Deng, A. (2015). Objective Bayesian two sample hypothesis testing for online controlled experiments. *WWW '15 companion: Proceedings of the 24th International Conference on World Wide Web, 24*, 923–928. https://doi.org/10.1145/2740908.2742563

Deng, A., & Lu, J., Chen, S. (2016). Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. *2016 IEEE International Conference on Data Science and Advanced Analytics, 3*, 243–252. https://doi.org/10.1109/DSAA.2016.33

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Goodson, M. (2014). *Most winning A/B test results are illusory*. Qubit. https://f.hubspotusercontent00.net/hubfs/215600/qubit-research-ab-test-results-are-illusory-1.pdf

Gronau, Q. F. (2019). *Abtest: Bayesian A/B testing*. R Foundation for Statistical Computing. https://CRAN.R-project.org/package=abtest

Gronau, Q. F., Raj, K. N. A., & Wagenmakers, E.-J. (2021). Informed Bayesian inference for the A/B test. *Journal of Statistical Software, 100*(17), 1–39. https://doi.org/10.18637/jss.v100.i17

Grünwald, P., de Heide, R., & Koolen, W. (2021). *Safe testing*. arXiv. https://doi.org/10.48550/arXiv.1906.07801

PsychOpen GOLD

Günel, E., & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika, 61*(3), 545–557. https://doi.org/10.1093/biomet/61.3.545

Hailpern, S. M., & Visintainer, P. F. (2003). Odds ratios and logistic regression: Further examples of their use and interpretation. *Stata Journal, 3*(3), 213–225. https://doi.org/10.1177/1536867X0300300301

Howard, J. V. (1998). The 2×2 table: A discussion from a Bayesian viewpoint. *Statistical Science, 13*(4), 351–367. https://www.jstor.org/stable/2676818

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods, 49*, 638–652. https://doi.org/10.3758/s13428-016-0739-8

JASP Team. (2020). *JASP* (Version 0.13) [Computer software]. JASP. https://jasp-stats.org/

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

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

Jennison, C., & Turnbull, B. W.(1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science, 5*(3), 299–317. https://www.jstor.org/stable/2245818

Johnson, G., Lewis, R. A., & Nubbemeyer, E. (2017). *The online display ad effectiveness funnel and carryover: Lessons from 432 field experiments*. SSRN. https://doi.org/10.2139/ssrn.2701578

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B, 72*(2), 143–170. https://doi.org/10.1111/j.1467-9868.2009.00730.x

Kamalbasha, S., & Eugster, M. J. (2021). *Bayesian A/B testing for business decisions*. In P. Haber, T. Lampoltshammer, M. Mayr, & K. Plankensteiner (Eds.), *Data science: Analytics and applications. Proceedings of the 3rd International Data Science Conference—iDSC2020* (pp. 50–57). Springer. https://doi.org/10.1007/978-3-658-32182-6_9

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B, 54*(1), 129–144. https://doi.org/10.1111/j.2517-6161.1992.tb01868.x

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience, 23*, 788–799. https://doi.org/10.1038/s41593-020-0660-4

King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(2), 171–184. https://doi.org/10.1586/erp.11.9

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

PsychOpen GOLD

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test., *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* (2nd ed.). Academic Press/Elsevier.

Kurt, W. (2019). *Bayesian statistics the fun way.* No Starch Press.

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

Lindley, D. V. (1972). *Bayesian statistics, a review.* SIAM.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15*(1), 22–25. https://doi.org/10.1111/j.1467-9639.1993.tb00252.x

Little, R. J. (1989). Testing the equality of two independent binomial proportions. *American Statistician, 43*(4), 283–288. https://doi.org/10.1080/00031305.1989.10475676

Ly, A., & Wagenmakers, E.-J. (2021). *Bayes factors for peri-null hypotheses* [Manuscript submitted for publication]. University of Amsterdam. https://arxiv.org/abs/2102.07162

McFarland, C. (2012). *Experiment! Website conversion rate optimization with A/B and multivariate testing*. New Riders.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. https://doi.org/10.1037/a0024377

Patel, N. (2018). *What is a good conversion rate? The answer might surprise you.* The Daily Egg. https://www.crazyegg.com/blog/what-is-good-conversion-rate/

Pham-Gia, T., Turkkan, N., & Eng, P. (1993). Bayesian analysis of the difference of two proportions. *Communications in statistics: Theory and methods, 22*(6), 1755–1771. https://doi.org/10.1080/03610929308831114

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (pp. 1-10). Scientific Research. http://www.R-project.org/conferences/DSC-2003/

Portman, F. (2017). *BayesAB: Fast Bayesian methods for A/B testing*. R Foundation for Statistical Computing. https://cran.r-project.org/web/packages/bayesAB/index.html

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have? Crisis and resolution in statistical inference. *American Statistician, 73*(3), 243–252. https://doi.org/10.1080/00031305.2017.1415971

Schmidt, M. N., & Mørup, M. (2019). Efficient computation for Bayesian comparison of two proportions. *Statistics & Probability Letters, 145*, 57–62. https://doi.org/10.1016/j.spl.2018.08.011

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods, 25*(2), 206–226. https://doi.org/10.1037/met0000234

PsychOpen GOLD

Stucchio, C. (2015). *Bayesian A/B testing at VWO*. Visual Website Optimizer. https://vwo.com/downloads/VWO_SmartStats_technical_whitepaper.pdf

Tang, S.-H., & Hall, V. C. (1995). The overjustification effect: A meta-analysis. *Applied Cognitive Psychology, 9*(5), 365–404. https://doi.org/10.1002/acp.2350090502

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine, 17*(8), 891–908. https://pubmed.ncbi.nlm.nih.gov/9595618/

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–8. https://doi.org/10.1177/2515245921992035

van Doorn, J., Matzke, D., & Wagenmakers, E.-J. (2020). An in-class demonstration of Bayesian inference. *Psychology Learning & Teaching, 19*(1), 36–45. https://doi.org/10.1177/1475725719848574

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & 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

von Frisch, K. (1914). *Der Farbensinn und Formensinn der Biene*. Gustav Fischer.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25*(3), 169–176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (2015). A power fallacy. *Behavior Research Methods, 47*(4), 913–917. https://doi.org/10.3758/s13428-014-0517-4

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*(2), 117–186. https://doi.org/10.1214/aoms/1177731118

Williams, M. N., Bååth, R. A., & Philipp, M. C. (2017). Using Bayes factors to test hypotheses in developmental research. *Research in Human Development, 14*(4), 321–337. https://doi.org/10.1080/15427609.2017.1370964

Wrinch, D., Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 42*(249), 369–390. https://doi.org/10.1080/14786442108633773

PsychOpen GOLD

Appendix

Key Statistical Results for the IBE Approach

Below we summarize three key results concerning the IBE approach to the Bayesian A/B test where the interest centers on the difference between two binomial chance parameters θ_1 and θ_2 that are assigned independent beta priors. Thus, we have $\theta_i \sim \text{Beta}(\alpha_i, \beta_i)$, $i = 1, 2$ with the interest on $\delta = \theta_1 - \theta_2$. As described in the main text, observing y_1 successes and $n_1 - y_1$ failures results in beta posterior distribution with parameters $\alpha_1 + y_1$ and $\beta_1 + n_1 - y_1$. To keep notation simple we assume that the updates have already been integrated into the parameters of the beta distribution.

The first result below gives an expression for the probability that $\theta_1 > \theta_2$ (Schmidt & Mørup, 2019); the second result gives an expression of the distribution for $\theta_1 - \theta_2$ (Pham-Gia et al., 1993); the third result shows how this beta difference distribution can be approximated by a normal distribution.

The Probability of $\theta_1 > \theta_2$

The posterior probability that $\theta_1 > \theta_2$ (Schmidt & Mørup, 2019) is given by

$$\Pr(\theta_1 > \theta_2) = \frac{Z(\alpha_1, \beta_1, \alpha_2, \beta_2)}{B(\alpha_1, \beta_1)B(\alpha_2, \beta_2)},$$

where $B(\alpha, \beta)$ is the Beta function and the normalizing constant Z is given by

$$Z(\alpha_1, \beta_1, \alpha_2, \beta_2) = \frac{\Gamma(\alpha_1 + \alpha_2)\Gamma(\beta_1 + \beta_2)}{\beta_1\alpha_2\Gamma(\alpha_1 + \beta_1 + \alpha_2 + \beta_2 - 1)}\, {}_3F_2\left[\begin{matrix}1, 1 - \alpha_1, 1 - \beta_2 \\ \beta_1 + 1, \alpha_2 + 1\end{matrix}; 1\right],$$

where ${}_3F_2$ is the generalized hypergeometric function.

The Distribution of $\delta = \theta_1 - \theta_2$

The distribution of $\delta = \theta_1 - \theta_2$ (Pham-Gia et al., 1993) has the following density:

For $0 < \delta \leq 1$,

$$f(\delta) = B(\alpha_2, \beta_1)\delta^{\beta_1 + \beta_2 - 1}(1 - \delta)^{\alpha_2 + \beta_1 - 1} \times F_1\left(\beta_1, \alpha_1 + \beta_1 + \alpha_2 + \beta_2 - 2, 1 - \alpha_1; \beta_1 + \alpha_2; (1 - \delta), 1 - \delta^2\right)/A,$$

and for $-1 \leq \delta < 0$

$$f(\delta) = B(\alpha_1, \beta_2)(-\delta)^{\beta_1 + \beta_2 - 1}(1 + \delta)^{\alpha_1 + \beta_2 - 1} \times F_1(\beta_2, 1 - \alpha_2, \alpha_1 + \alpha_2 + \beta_1 + \beta_2 - 2; \alpha_1 + \beta_2; 1 - \delta^2, 1 + \delta)/A,$$

where $A = B(\alpha_1, \beta_1)B(\alpha_2, \beta_2)$ and F_1 is Appell's first hypergeometric function.

Moreover, if $\alpha_1 + \alpha_2 > 1$ and $\beta_1 + \beta_2 > 1$ we have:

$f(\delta = 0) = B(\alpha_1 + \alpha_2 - 1, \beta_1 + \beta_2 - 1)/A$

PsychOpen GOLD

The Normal Approximation to the Difference Between Two Beta Distributions

The *Beta*(α, β) distribution can be approximated by the normal distribution when α and β are sufficiently large:

$$\text{Beta}(\alpha, \beta) \;\dot{\sim}\; \text{Normal}\left(\mu = \frac{\alpha}{\alpha + \beta}, \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}\right).$$

Also, the distribution of the difference between two normally distributed variables $X \sim N(\mu_X, \sigma^2{}_X)$ and $Y \sim N(\mu_Y, \sigma^2{}_Y)$ is given by another normal distribution as $X - Y \sim N(\mu_X - \mu_Y, \sigma^2{}_X + \sigma^2{}_Y)$. When the parameters of the composite beta distributions are sufficiently large, the difference between a *Beta*(α_2, β_2) and a *Beta*(α_1, β_1) distribution is therefore approximated as follows:

$$\text{Beta}(\alpha_2, \beta_2) - \text{Beta}(\alpha_1, \beta_1) \;\dot{\sim}\; \text{Normal}\left(\mu = \frac{\alpha_2}{\alpha_2 + \beta_2} - \frac{\alpha_1}{\alpha_1 + \beta_1}, \sigma^2 = \frac{\alpha_2\beta_2}{(\alpha_2 + \beta_2)^2(\alpha_2 + \beta_2 + 1)} + \frac{\alpha_1\beta_1}{(\alpha_1 + \beta_1)^2(\alpha_1 + \beta_1 + 1)}\right).$$

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.

PsychOpen GOLD