Check for updates

METHOD ARTICLE

Using Bayesian test relationship between categorical random variables on small sample sizes: application on the relationship between the weight status of parents and their children in understudied populations living in Pacific Island countries and territories

[version 1; peer review: 2 approved with reservations]

Guillaume Wattelez[1], Thibaut Demaneuf[2], Solène Bertrand-Protat[2], Olivier Galy[1]

[1]Institute for Societies, Languages and Education, University of New Caledonia, Noumea, South Province, 98800, New Caledonia
[2]Public Health Division, Pacific Community, Noumea, South Province, 98800, New Caledonia

v1 **First published:** 22 Aug 2026, **6**:301
https://doi.org/10.12688/openreseurope.24052.1
Latest published: 22 Aug 2026, **6**:301
https://doi.org/10.12688/openreseurope.24052.1

Abstract

Background

Within the traditional frequentist framework, the emphasis is often on reporting the p-value. Obtaining significant results is infrequent in small sample size contexts, implying a major concern when reporting results in scientific research because rejecting the null hypothesis does not provide evidence of the absence of an effect.

Methods

We explored the association between child weight status and parental weight status within an understudied Pacific Island population. The sample was divided in two groups: Group 1 with children whose parents were not both overweight, and Group 2 containing children whose both parents were overweight. We compared:

- the Fisher's exact test and computed its p-value;
-

Open Peer Review

Approval Status ? ?

	1	2
version 1 22 Aug 2026	? view	? view

1. **Julien Jean Malard-Adam**, Universite de Montpellier, Montpellier, France
2. **Valen E. Johnson**, Texas A&M University, College Station, USA

Any reports and responses or comments on the article can be found at the end of the article.

the Bayesian A/B test with a variety of indicators including the Bayes factors (BF), posterior probabilities, and density functions.

Two Bayesian approaches were compared in order to evaluate the hypothesis that the probability of success was higher in Group 2 than in Group 1 (H+): the Independent Beta Estimation (IBE) and the Logit Transformation Testing (LTT). Furthermore, the data were slightly modified and the indicators were re-examined to understand the effects of such minor alterations on the results and conclusions derived from small samples sizes.

Results

- Estimates of the success probabilities: 0.17 in Group 1, 0.52 in Group 2.
- Fisher's exact test: not significant ($p \approx 0.182$) with a large effect size associated ($h=0.78$).
- Evidence for H+ with IBE: moderate to strong (BF from 8.5 to 14.9).
- Evidence for H+ with LTT: negligible to moderate (BF from 2.2 to 13.9).
- Adding a single observation: conclusions not changed for both the Fisher's exact test and the Bayesian approaches.
- Doubling the sample size: significant result for the Fisher's exact test and the evidence for H+ strengthened.

Conclusions

The Bayesian testing framework offers additional insights that enrich interpretation alongside frequentist testing. Scientific studies would benefit from giving them greater consideration.

Plain Language summary

Small samples pose persistent challenges for statistical inference in scientific research, especially in applied health research. Within the traditional frequentist framework, emphasis is often placed on the p-value; yet in small-sample settings, statistically significant results are uncommon, and failure to reject the null hypothesis does not

constitute evidence of no effect. Conversely, employing statistical indicators from the Bayesian testing framework can offer additional insights that are valuable for further discussion of the results. This paper demonstrates how Bayesian A/B testing can complement conventional analyses through a case study aimed at exploring the association between child weight status and parental weight status in an understudied Pacific Island population. We compare the Independent Beta Estimation and Logit Transformation Testing approaches, examining a range of indicators such as Bayes factors, posterior probabilities, credible intervals, and posterior densities. We also investigate the effect of slight data alterations in order to show how sensitive conclusions may be when samples are small. Adding a single observation did not substantially change the conclusions for both the Fisher's exact test and the Bayesian approaches, although the values of the Bayes factors may have been affected. However, doubling the sample size made the Fisher's exact test significant and strengthened the primary conclusions of the Bayesian tests. Taken together, the results illustrate how Bayesian indicators can enrich interpretation alongside frequentist testing and help distinguish absence of evidence from evidence of absence.

Keywords
Bayesian A/B testing, Bayes factor, frequentist testing, Fisher's exact test, small-sample inference, childhood obesity, public health, Oceania



This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.



This article is included in the Horizon 2020 gateway.



This article is included in the Family Farming, Lifestyle and Health in Small Islands, Countries and Territories collection.

Corresponding author: Guillaume Wattelez (guillaume.wattelez@unc.nc)

Author roles: Wattelez G: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Demaneuf T**: Writing – Original Draft Preparation, Writing – Review & Editing; **Bertrand-Protat S**: Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Galy O**: Funding Acquisition, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The study related to the Move Ô program was funded by the Pacific Fund (MAE France) and the Pacific Community (SPC). The project related to the Lifou case study has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 873185: (Family farming, lifestyle, and health in the Pacific [FALAH]); and the Pacific Island University Research Network grant 2016 N° 042016. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Wattelez G *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Wattelez G, Demaneuf T, Bertrand-Protat S and Galy O. **Using Bayesian test relationship between categorical random variables on small sample sizes: application on the relationship between the weight status of parents and their children in understudied populations living in Pacific Island countries and territories [version 1; peer review: 2 approved with reservations]** Open Research Europe 2026, **6**:301 https://doi.org/10.12688/openreseurope.24052.1

First published: 22 Aug 2026, **6**:301 https://doi.org/10.12688/openreseurope.24052.1

Introduction

Researchers often seek to clarify relationships between two categorical variables. For example, statistical tests are commonly used to assess whether two proportions differ across two sampled populations. In public health and the social sciences, such comparisons frequently concern sex, gender, ethnic group, socioeconomic status, or behavioural categories.[1–4] The conventional approach is the frequentist framework. In practice, this involves setting a Type I error threshold, usually at 5%, selecting an appropriate test, and comparing the resulting p-value with that pre-specified threshold.[5] Additional indicators, such as effect size and statistical power, are sometimes reported as well.[6,7] By design, frequentist tests focus on a test statistic whose distribution is known under the null hypothesis of no effect. The p-value therefore corresponds to the probability of observing data at least as extreme as those obtained, assuming that the null hypothesis is true.[8]

However, as highlighted by Dablander et al. (2022) and Keysers et al. (2020), frequentist hypothesis tests can provide evidence for an association when the null hypothesis is rejected, but when it is not rejected, they indicate only an absence of evidence.[8,9] As discussed by Altman and Bland (1995), failure to reject the null hypothesis should therefore not be interpreted as confirmation that no association exists.[10] This situation is particularly common in studies based on small sample sizes, where limited statistical power reduces the ability to detect true effects.[8]

Although minimum sample sizes can be estimated from assumptions about power and effect size,[6,11] logistical or resource constraints may prevent researchers from reaching the recommended sample size.[12] Geographic remoteness, scarcity, and population dispersion make some populations particularly difficult to study. Research conducted in Pacific Island Countries and Territories (PICTs) exemplifies many of the structural constraints associated with small-sample research. Geographic isolation, population dispersion across multiple islands, and limited local research infrastructure often restrict achievable sample sizes.[13] In addition, individual islands may have only a few thousand inhabitants.[14,15] In this contexts, conventional hypothesis testing frequently produces non-significant results, not necessarily because associations are absent, but because the available data do not contain enough information to meet conventional decision thresholds.

This situation is especially problematic in public health research, where small populations are often those experiencing a disproportionate burden of disease and where evidence is most urgently needed to inform policy and intervention. Indeed, PICTs face a rapidly increasing burden of non-communicable diseases, including overweight and obesity in both adults and children.[16,17] The combination of high public health relevance and limited statistical power underscores the need for analytical frameworks that can extract meaningful information from small datasets.

As an example, Bertrand-Protat et al. (2024) presented data on an understudied child population in Wallis following a school-based physical activity promotion programme.[18] Of the 52 pupils who consented to participate, only 32 participants with complete data were included in the reported analyses. Although the authors conducted frequentist tests, they were able to report only descriptive results without statistical significance.[18] Those results nevertheless did not justify concluding that the data supported the null hypothesis, especially given the particularly small sample size.[8,10] This example illustrates the critical need to reconsider how evidence from small-sample studies is analysed and interpreted in Pacific research, and more broadly in social sciences.[19] Methodological approaches that allow a more nuanced interpretation of results obtained from small samples are therefore essential to ensure that research conducted in such settings is not systematically undervalued.

For its part, the Bayesian testing framework offers additional indicators, such as weighted evidence for or against the hypotheses under consideration, derived from Bayes factors.[8,20–22] Examination of posterior densities and the credible intervals (CI) for the relevant parameters also provides valuable information. Unlike the frequentist framework, which relies solely on assumptions related to the null hypothesis, the Bayesian approach also requires an explicit formulation of the alternative hypothesis and an updating step based on that formulation and the observed data.[8] In small-sample contexts, this distinction is particularly important because non-significant p-values are common and are often misinterpreted. Furthermore, Bayesian methods make it possible to distinguish between absence of evidence and evidence of absence.[10] By presenting both frequentist and Bayesian results side by side, this study illustrates how the two frameworks can jointly support a more informative and transparent interpretation of empirical findings.

This study provides an overview of Bayesian A/B testing, with a particular emphasis on the interpretation of its outputs, thereby demonstrating its usefulness for researchers comparing two proportions in small-sample settings. We implement both the Independent Beta Estimation (IBE) and the Logit Transformation Testing (LTT) approaches using several R-based implementations or computational strategies (IBE: 3; LTT: 2). These approaches will be applied to data from an empirical study conducted in a PICT in order to examine the association between parental weight status and child weight

status. Because small samples can be highly sensitive to minor changes in the data, we also investigate the effect of slight data alterations on the results.

The Materials and methods section first describes the case study context. It then briefly reviews the frequentist framework - specifically the Fisher's exact test, which is commonly used in such situations - and the Bayesian framework, focusing on the IBE and LTT approaches. The priors selected for the Bayesian analyses are then detailed, and the section concludes with a description of the Bayes factor calculations. The results are subsequently presented and finally discussed.

Material and methods

Move Ô study: context and dataset

The Move Ô protocol was developed to generate actionable evidence for policymakers addressing childhood obesity in Pacific Island contexts. Its primary objective was to examine a broad range of determinants of childhood overweight and obesity, including factors that are well established internationally - such as unhealthy dietary patterns, physical inactivity and sedentary behaviour, insufficient sleep, parental body weight, and socioeconomic conditions - but that have so far been insufficiently studied in the Pacific region.[18] The study was conducted among all Grade 5 students (CM2 in French school system) from four schools on Wallis Island. A total of 52 children aged 9 to 11 years participated in the study in March 2024. Data were collected using multiple complementary methods, including questionnaires, wrist-worn accelerometers, interviews, and direct observations. The study received approval from the New Caledonia Ethics Advisory Committee and complied with the Declaration of Helsinki.

Children's weight was measured using a SECA 813 scale, and their height was measured with a SECA 213 stadiometer. Parental height and weight were collected by questionnaire, which asked: "Please provide the following information about the child's father (respectively, the child's mother): height: … ; weight: …". Only participants with complete height and weight data for both the child and the parents were included in the present analysis, yielding a final sample of 27 participants.

Weight status categorisation

Body mass index (BMI) for both children and adults was calculated from weight in kilograms and height in metres as follows:

$$BMI = \frac{weight}{height^2} \tag{1}$$

Adults were classified as either "Underweight or Normal range" or "Overweight or Obese" in accordance with World Health Organization guidelines[23]: "Underweight or Normal range" if $BMI < 25$, and "Overweight or Obese" if $BMI \geq 25$. Children were classified using the International Obesity Task Force reference.[24] As for adults, the children were grouped into two categories: "Underweight or Normal range" and "Overweight or Obese".

Definitions and notations

Group 1 comprised children with at least one parent in the "Underweight or Normal range" category, whereas Group 2 comprised children whose parents were both "Overweight or Obese".

We denote:

- n_1 the total number of observations and y_1 the number of successes (i.e. the number of overweight or obese children) in Group 1;
- n_2 the total number of observations and y_2 the number of successes (i.e. the number of overweight or obese children) in Group 2;
- π_1 the true proportion of successes among children in the population whose parents are not both overweight or obese (i.e. the population with characteristics similar to Group 1);
- π_2 the true proportion of successes among children in the population whose parents are both overweight or obese (i.e. the population with characteristics similar to Group 2);
- $\widehat{\pi}_i$, $i \in \{1,2\}$, the sample estimates of the true success proportion π_i.

In hypothesis testing, the plausibility of two competing propositions is compared. The null hypothesis assumes equal proportions. The four alternative hypotheses considered are as follows:

- $H_0 : \pi_1 = \pi_2$
- $H_1 : \pi_1 \neq \pi_2$
- $H_- : \pi_1 > \pi_2$
- $H_+ : \pi_1 < \pi_2$

When comparing two proportions, it is also common to report an effect size,[6] denoted h, defined as follows:

$$h = 2 \times |\sin^{-1}\sqrt{\widehat{\pi}_2} - \sin^{-1}\sqrt{\widehat{\pi}_1}| \tag{2}$$

Reminder about the frequentist methods and the Fisher's exact test

Frequentist hypothesis testing assesses whether the observed data is statistically compatible with the null hypothesis of "no effect".[5] When comparing two proportions, this approach typically involves testing the null hypothesis of equality between groups against an alternative hypothesis specifying a difference. Researchers are primarily interested in the p-value, i.e. the probability of observing data as extreme as, or more extreme than, those observed, assuming that the null hypothesis is true. This p-value is then compared with a pre-specified significance threshold, which is most commonly set at 0.05. Although widely used, this framework does not provide a direct measure of evidence in favour of the null hypothesis and can be difficult to interpret in small-sample situations.

Fisher's exact test is a non-parametric statistical test commonly used to assess the association between two categorical variables in contingency tables, particularly when small sample sizes make the standard χ^2 large-sample test unreliable. Based on the hypergeometric distribution, this test computes the exact probability of observing a difference at least as extreme as the one obtained under the null hypothesis, i.e. a difference that can be explained by chance.[25] In the present study, Fisher's exact test is applied in both two-sided and one-sided forms to compare the proportions of overweight or obese children between the two parental weight-status groups. As with other frequentist tests, failure to reject the null hypothesis should be interpreted as an absence of evidence rather than evidence of no association.

Principles of the Bayesian A/B testing models

A binomial model is used to describe the number of overweight or obese children in each group, i.e. y_1 in Group 1 and y_2 in Group 2:

$$\begin{aligned} y_1 &\sim \mathcal{B}(n_1, \pi_1) \\ y_2 &\sim \mathcal{B}(n_2, \pi_2) \end{aligned} \tag{3}$$

Where $\mathcal{B}$ denotes the binomial distribution and π_i, $i \in \{1,2\}$, represents the underlying probability of success in group i. The aim is to infer plausible values for π_i, $i \in \{1,2\}$ and to quantify the evidence supporting competing hypotheses about their relationship.

In the Bayesian approach, the analysis begins with prior distributions, which reflect the experimenter's beliefs or knowledge about the parameters of interest (i.e. π_i) before data collection.[19,26] Priors may be non-informative, reflecting limited prior knowledge, or informative, reflecting external evidence or strong expectations about the parameters.[3] Once the data have been collected, each prior is updated to a posterior distribution (Figure 1), which then helps determine which hypothesis is better supported.

Within the Bayesian framework, the strength of evidence favouring one hypothesis over another is quantified by the posterior odds. This quantity is related to the Bayes factor, as shown in the following equation used to compare H_+ and H_0.[21,27]

$$\frac{P(H_+|data)}{P(H_0|\,data)} = BF_{+0} \times \frac{P(H_+)}{P(H_0)} \tag{4}$$

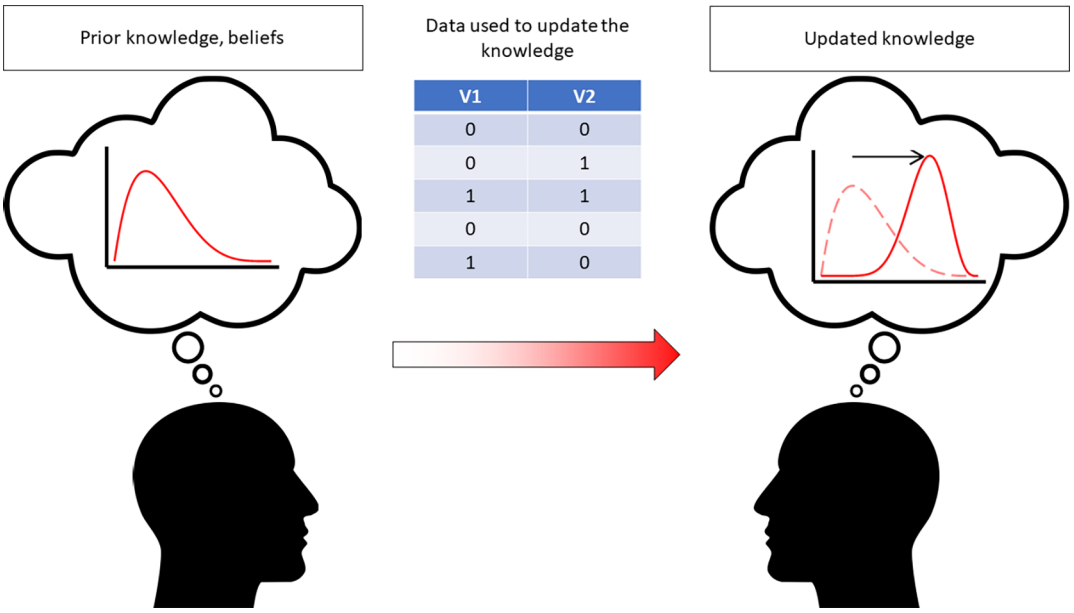


Figure 1. Diagram illustrating the principles underlying the Bayesian tests.

In this equation, $P(H_+|\,data)$ and $P(H_0|data)$ are the posterior probabilities of H_+ and H_0, respectively, whereas $P(H_+)$ and $P(H_0)$ denote the prior probabilities, based on beliefs and knowledge, of H_+ and H_0, respectively. The Bayes factor is denoted as $BF_{+0} = \frac{P(data|H_+)}{P(data|H_0)}$.

The two approaches considered in this paper, namely IBE and LTT, are based on distinct models and therefore on different families of prior distributions, which in turn lead to different posterior distributions.

Independent Beta Estimation (IBE) approach

Although readers can find a detailed description of the IBE approach in Hoffmann et al. (2022),[3] we reiterate the main points here.

In the IBE approach, the distributions of y_1 and y_2 are assumed to be independent, just as the success probabilities are assigned independent βeta distributions: $\pi_i \sim \beta eta(\alpha_i, \beta_i)$, $i \in \{1,2\}$. These distributions model the relative plausibility for π_i, $i \in \{1,2\}$, with the α_i and β_i parameters interpreted as counts of hypothetical successes and failures, respectively.

At the outset, prior distributions, denoted as $f(\pi_i)$, are selected on the basis of the experimenter's knowledge, or lack thereof, and therefore be informative or non-informative. The data collected during the experiment are then used to update the prior distributions into posterior distributions, according to Bayes' rule for $i \in \{1,2\}$:

$$P(\pi_i|data) = \frac{P(\pi_i) \times P(data|\,\pi_i)}{P(data)} \tag{5}$$

where *data* represents knowledge of the number of observations (i.e. n_i) and the number of successes (i.e. y_i) derived from the observed data; $P(\pi_i)$ is the probability deduced from $f(\pi_i)$; $P(data)$ arises from the binomial distribution of the data; and $P(data|\pi_i)$ is the likelihood of the data given the parameter π_i.

In this approach, since the prior distributions are βeta distributions with parameters (α_i, β_i), the posterior distributions become βeta distributions with parameters $(\alpha_i + s_i, \beta_i + f_i)$, where s_i is the number of observed successes and f_i the number of observed failures.[28] However, as pointed by Gronau et al. (2021) and Dablander et al. (2022), this test does not allow the experimenter to assess exact equality of the proportions.[9,29] In practice, it therefore amounts to comparing H_+ and H_-.

To characterise prior and posterior uncertainty, we rely on simulation from βeta distributions using standard R functions,[30,31] together with results obtained from the *bayesAB* package and from Markov chain Monte-Carlo

(MCMC) estimation via the *R2jags* package in order to verify consistency across computational strategies.[19,30,32–36] R version 4.4.2 was used in the present study.

In the results, we consider the computed and simulated distributions of the parameters of interest, i.e. π_1 and π_2. Given its great interest, we also examine the assessment of the difference between the proportions, denoted $\delta = \pi_2 - \pi_1$, and estimate the 90% highest density interval (HDI) and equal-tailed interval (ETI) for the credible interval (CI),[37] the posterior probability $P(H_+|data) = P(\pi_1 < \pi_2|data)$, and the Bayes factors $BF_{+-} = \frac{P(data|H_+)}{P(data|H_-)}$.

Logit Transformation Testing (LTT) approach

As in the previous subsection, readers can refer to Hoffmann et al. (2022) for a comprehensive description of the LTT approach.[3] However, we summarise the main elements hereafter.

The LTT approach is based on a model described by Kass and Vaidyanathan (1992),[38] who proposed a test applied to the log-odds transformation of the parameters of interest, namely π_1 and π_2. The model is thus expressed as follows:

$$\begin{aligned} y_1 &\sim \mathcal{B}(n_1, \pi_1) \\ y_2 &\sim \mathcal{B}(n_2, \pi_2) \\ \log\left(\frac{\pi_1}{1-\pi_1}\right) &= \gamma - \frac{\psi}{2} \\ \log\left(\frac{\pi_2}{1-\pi_2}\right) &= \gamma + \frac{\psi}{2} \end{aligned} \tag{6}$$

where γ denotes the grand mean of the odds ratio and ψ represents the distance between the two conditions.[39] By default, both parameters are assumed to follow Gaussian distributions: $\gamma \sim \mathcal{N}\left(\mu_\gamma, \sigma_\gamma^2\right)$ and $\psi \sim \mathcal{N}\left(\mu_\psi, \sigma_\psi^2\right)$. Within this framework, the hypotheses to be tested can be written as follows:

- $H_0 : \psi = 0$
- $H_1 : \psi \neq 0$
- $H_- : \psi < 0$
- $H_+ : \psi > 0$

Updating the (γ, ψ) parameters then yields posterior distributions, modal estimates for the parameters of interest (π_1, π_2), and Bayes factors for hypothesis comparison.

Results from the implementation included in the *abtest* R package are presented in this paper.[29] In addition, we incorporate information from the *R2jags* outputs used in the LTT approach.[33] In that implementation, however, we chose to not use truncated distributions, unlike the *abtest* R package. In the results, we consider the outputs from *abtest*, which provide information on Bayes factors ($BF_{+-} = \frac{P(data|H_+)}{P(data|H_-)}$ and $BF_{+0} = \frac{P(data|H_+)}{P(data|H_0)}$), posterior probabilities ($P(H_+|data) = P(\pi_1 < \pi_2|data)$ versus $P(H_-|data) = P(\pi_1 > \pi_2|data)$ or $P(H_0|data) = P(\pi_1 = \pi_2|data)$), 90% CI, and robustness analysis.[29] We also examine the distributions of the parameters of interests and the difference δ.

Choices of prior distributions

To reflect realistic analytical practice and recommendations in the methodological literature, we examine multiple prior scenarios rather than relying on a single prior specification.[3,19,31] This is particularly important in our small-sample settings, where inferential summaries - especially Bayes factors - can be sensitive to prior choices.[8] Accordingly, we investigate three scenarios: non-informative, informative, and optimistic. These scenarios are intended to span a reasonable range of prior knowledge and expectations relevant to childhood overweight and obesity in Pacific settings.

Before describing the priors used in the three scenarios (non-informative, informative, and optimistic), it is worth noting that, in the LTT approach, the central values of both γ and ψ can be derived from the beliefs and knowledge available - or unavailable - about the primary parameters (i.e. π_1 and π_2), according to the following relationships.

$$\begin{cases} \log\left(\frac{\pi_1}{1-\pi_1}\right) = \gamma - \frac{\psi}{2} \\ \log\left(\frac{\pi_2}{1-\pi_2}\right) = \gamma + \frac{\psi}{2} \end{cases} \Leftrightarrow \begin{cases} \gamma = \frac{1}{2} \times \left(\log\frac{\pi_1}{1-\pi_1} + \log\frac{\pi_2}{1-\pi_2}\right) \\ \psi = \log\frac{\pi_2}{1-\pi_2} - \log\frac{\pi_1}{1-\pi_1} \end{cases} \quad (7)$$

For the **non-informative scenario** in the IBE approach, every value between 0 and 1 is, a priori, equally possible for the success proportion (i.e. overweight or obese) in each of the two groups under study. Two $\beta eta(1,1)$ distributions therefore provide an appropriate prior specification, which is equivalent to using two uniform distributions defined on $[0,1]$ (Figure 2). In the LTT approach, the absence of prior knowledge can be modelled by values of π_1 and π_2 distributed around 0.5. Consequently, choosing two $\mathcal{N}(0,1)$ distributions for both γ and ψ is appropriate for this prior (Figure 2).

When selecting an **informative prior distribution**, the experimenter must draw on prior beliefs and knowledge.[26,40] Recent studies suggest that approximately 20% to 30% of young people in the considered populations are affected by this health issue.[17,41] This information guides the choice of the parameters of the βeta distribution so that the density peaks around these values. For example, using two $\beta eta(2,5)$ distributions in the IBE approach, with substantial mass around 0.2, is consistent with previous observations (Figure 2). In the LTT approach, if the researcher does not distinguish Groups 1 and 2, the underlying belief is that $\pi_1 = \pi_2 = 0.2$. From Equation 7, we deduce:

$$\begin{cases} \gamma = \frac{1}{2} \times \left(\log\frac{0.2}{1-0.2} + \log\frac{0.2}{1-0.2}\right) \approx -1.4 \\ \psi = \log\frac{0.2}{1-0.2} - \log\frac{0.2}{1-0.2} = 0 \end{cases} \quad (8)$$

This is why using a Gaussian distribution centred at -1.4 for γ and another centred at 0 for ψ appears reasonable. The standard deviation can be set to 0.5 in order to limit dispersion, yielding the following prior distributions (Figure 2).

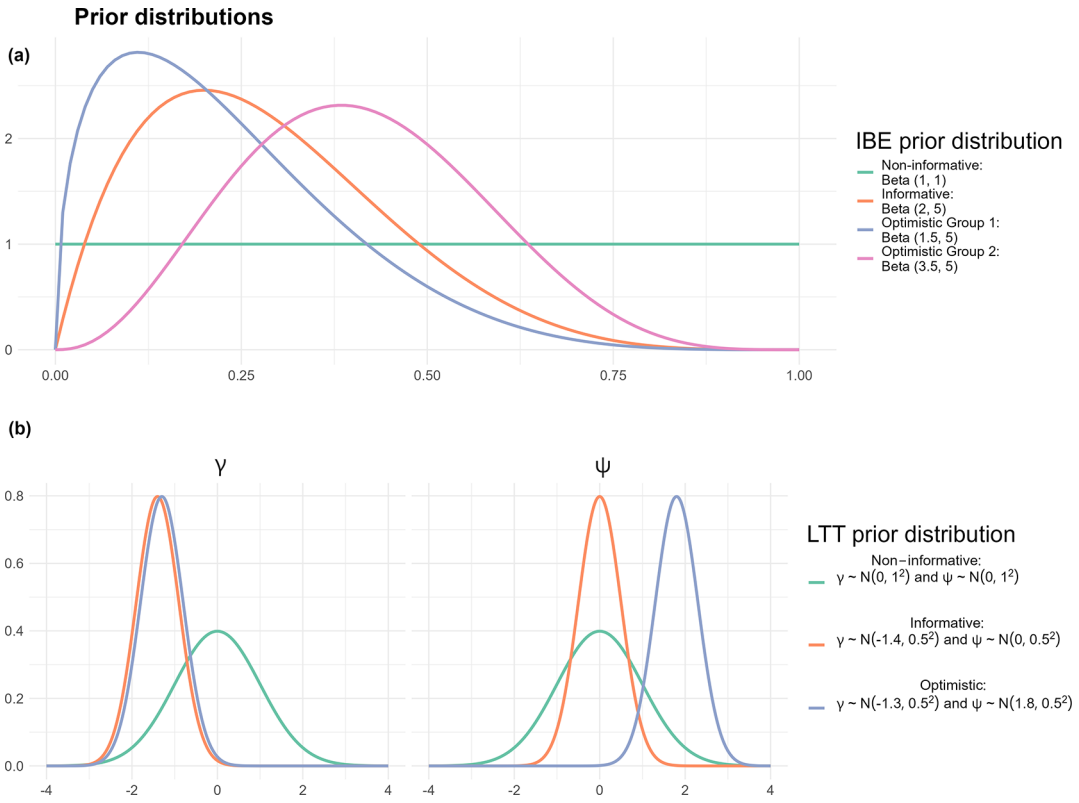


Figure 2. Prior distributions used in the present study for each scenario investigated (1. Non-informative, 2. Informative, 3. Optimistic): (a) βeta distributions used to estimate the π_1 and π_2 parameters in the IBE approach, (b) Normal distributions used to estimate the γ and ψ parameters in the LTT approach.

$$\begin{aligned}\gamma &\sim \mathcal{N}\left(-1.4, 0.5^2\right)\\ \psi &\sim \mathcal{N}\left(0, 0.5^2\right)\end{aligned} \tag{9}$$

For the **optimistic scenario**, the experimenter again relies on prior beliefs and knowledge.[26,40] For genetic or environmental reasons, it is reasonable to expect that a child in Group 2 (whose parents are both overweight or obese) is much more likely to be overweight than a child in Group 1 (with at most one parent who is overweight or obese).[42–44] In the IBE approach, this scenario is modelled by two priors depending on the group. The experimenter could therefore select a $\beta eta(1.5,5)$ distribution, with substantial mass around 0.1, for Group 1, and a $\beta eta(3.5,5)$ distribution, with substantial mass around 0.4, for Group 2 (Figure 2).

From Equation 7, with $\pi_1 = 0.1$ and $\pi_2 = 0.4$, we obtain in the LTT approach:

$$\begin{cases} \gamma = \dfrac{1}{2} \times \left(\log \dfrac{0.1}{1-0.1} + \log \dfrac{0.4}{1-0.4} \right) \approx -1.3 \\ \psi = \log \dfrac{0.4}{1-0.4} - \log \dfrac{0.1}{1-0.1} = 1.8 \end{cases} \tag{10}$$

Consequently, using a Gaussian distribution centred at -1.3 for γ and another centred at 1.8 for ψ appears reasonable. As in the informative scenario, the standard deviation can be set to 0.5 in order to limit dispersion, yielding the following prior distributions (Figure 2).

$$\begin{aligned}\gamma &\sim \mathcal{N}\left(-1.3, 0.5^2\right)\\ \psi &\sim \mathcal{N}\left(1.8, 0.5^2\right)\end{aligned} \tag{11}$$

Calculation of Bayes factors

In the IBE approach, it is possible to compare the evidence supporting either H_+ or H_-.[9] The resulting Bayes factor, BF_{+-}, was computed as follows:

$$BF_{+-} = \frac{P(H_+|data)}{P(H_-|\,data)} = \times \frac{P(H_-)}{P(H_+)} \tag{12}$$

where $P(H_+|\,data)$ and $P(H_-|data)$ are the posterior probabilities respectively for H_+ and H_- respectively, and $P(H_+)$ and $P(H_-)$ are the prior probabilities, estimated from the number of simulated values supporting H_+ and H_-, respectively, under the prior βeta distributions. Similarly, the posterior probabilities were estimated by the number of simulated values supporting H_+ or H_- under the posterior βeta distributions.

In the LTT approach, the *ab_test* function from *abtest* R package provides estimates of Bayes factors in favour of H_1 over H_0, H_+ over H_- and H_1 over H_0.[29] When using the MCMC simulations generated with the *R2jags* package, the Savage-Dickey method was used to estimate BF_{01}, from which $BF_{10} = 1/BF_{01}$ was then derived.[31,45] The remaining Bayes factors, i.e. BF_{+0} and BF_{-0}, were subsequently calculated from the estimates of BF_{10} and from the prior and posterior probabilities[46]: $BF_{+0} = BF_{10} \times \frac{P(H_+|data)}{P(H_+)}$ and $BF_{-0} = BF_{10} \times \frac{P(H_-|data)}{P(H_-)}$.

When comparing H_+ and H_-, the Bayes factor was estimated as $BF_{+-} = BF_{+0}/BF_{-0}$, and the posterior probability was estimated as $P(H_+|data) = \frac{BF_{+-} \times P(H_+)}{BF_{+-} \times P(H_+) + P(H_-)}$.

Application results

Results of the Fisher's exact test for the Move Ô case study

Table 1 presents the contingency table summarising the distribution of the successes across the two groups. From this table, the estimates of y_1 and y_2 are $\widehat{y}_1 = 1$ and $\widehat{y}_2 = 11$, respectively. Likewise the estimates of π_1 and π_2 are $\widehat{\pi}_1 = 0.17$ and $\widehat{\pi}_2 = 0.52$, respectively, yielding $h = 0.78$, which corresponds to a large effect size.[6]

Applying the two-sided Fisher's exact test, i.e. testing H_0 against H_1, yielded $p \approx 0.182$; therefore H_0 was not rejected. Under the frequentist framework, the data therefore provide no evidence of a significant difference between the two groups. Applying the one-sided Fisher's exact test, i.e. testing H_0 against H_+, yielded $p \approx 0.139$; again H_0 was not rejected. As in the two-sided analysis, the data therefore provide no evidence of a significant difference between the two groups.

Results of the IBE approach for the Move Ô case study

We now turn to the IBE and LTT approaches. Under the IBE approach, the posterior distributions can be derived directly from the prior specification and the observed data. Table 2 summarises the prior and posterior distributions for Groups 1 and 2 under the three scenarios investigated, namely the non-informative, informative, and optimistic scenarios. Figure 3 shows the shape of each distribution, and Figure 4 presents the distributions of the difference $\delta = \pi_2 - \pi_1$ simulated under both the prior and posterior settings.

Figure 3 clearly shows how the density curves change once the sample information is incorporated. In the non-informative scenario, the flat prior distributions become posterior distributions with pronounced peaks. Overall, the choice of scenario (non-informative, informative, and optimistic) did not substantially alter the general shape of the posterior distributions or their main indicators. In the non-informative scenario, the posterior density peaked at approximately 0.18 in Group 1 and 0.52 in Group 2. In the informative scenario, the corresponding modes were approximately 0.19 and 0.46. In the optimistic scenario, they were approximately 0.15 for Group 1 and 0.49 for Group 2. In all scenarios, the support of the density was wider for Group 1 than for Group 2 because the sample size was smaller in Group 1.

Figure 4 shows that, regardless of the method or the package used, the posterior distributions of δ were broadly similar, although the posterior distribution was more diffuse in the non-informative scenario. In the non-informative scenario, the prior distribution peaked around 0 and the posterior distribution around 0.32. In the informative scenario, the prior again

Table 1. Distribution of successes (overweight or obese) and failures (underweight or normal) for each Group (1: parents are not overweight or obese together; 2: both parents are overweight or obese) for the Move Ô case study.

	Group 1 [n (%)]	Group 2 [n (%)]
Underweight or normal	5 (83)	10 (48)
Overweight or obese	1 (17)	11 (52)

Table 2. Prior and posterior distributions according to the investigated scenario (1. Non-informative, 2. Informative, 3. Optimistic) and the Group in the IBE approach applied on the Move Ô case study.

	Prior distribution		Posterior distribution	
	Group 1	Group 2	Group 1	Group 2
Non-informative	$\beta eta(1,1)$	$\beta eta(1,1)$	$\beta eta(2,6)$	$\beta eta(12,11)$
Informative	$\beta eta(2,5)$	$\beta eta(2,5)$	$\beta eta(3,10)$	$\beta eta(13,15)$
Optimistic	$\beta eta(1.5,5)$	$\beta eta(3.5,5)$	$\beta eta(2.5,10)$	$\beta eta(14.5,15)$

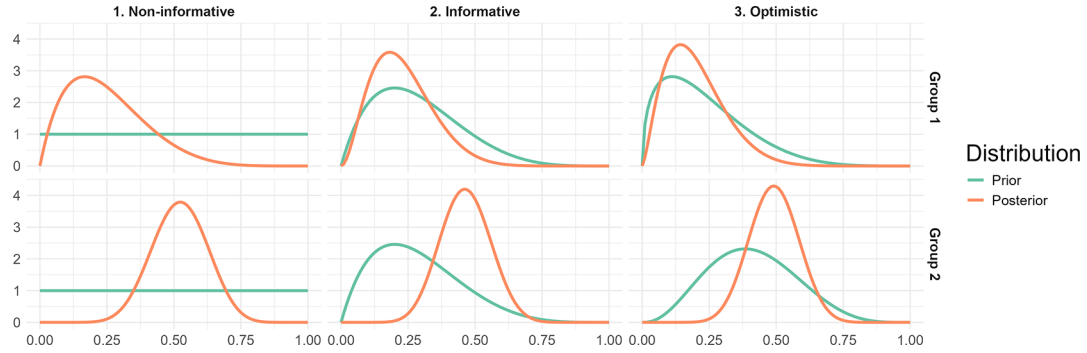


Figure 3. Prior and posterior βeta distributions of the IBE approach applied on Groups 1 (upper row) and 2 (lower row) for the Move Ô case study for each scenario investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

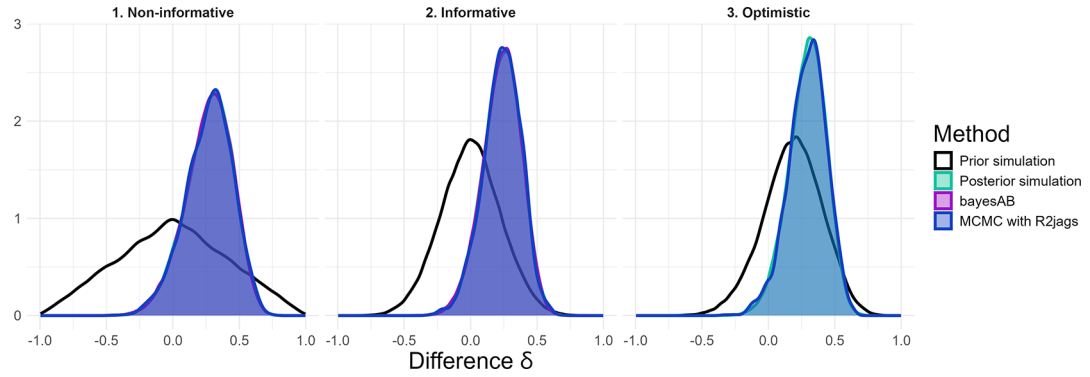


Figure 4. Distributions of the difference ($\delta = \pi_2 - \pi_1$) simulated in the prior and posterior settings for the Move Ô case study by the three implementations (our own simulation coding, bayesAB package, and MCMC with R2jags) of the IBE approach for each scenario investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

peaked around 0, whereas the posterior peaked around 0.28. In the optimistic scenario, the prior peaked around 0.21 and the posterior around 0.31.

Table 3 reports the posterior probabilities $P(H_+ \mid data)$ and the Bayes factors BF_{+-}, allowing the comparison of the three implementations of the IBE approach across the non-informative, informative, and optimistic scenarios. Table S1 complements these results by reporting the quantiles and the 90% CI of the parameter distributions (π_1, π_2, and δ). Across all scenarios, the three methods yielded equivalent highly similar results. For example, the only posterior quantiles that differed at the second decimal place were approximately the 0% and 100% quantiles (Table S1). The posterior probabilities $P(H_+|data)$ showed only minor variations for the MCMC implementation with *R2jags*, at the second decimal place in the optimistic scenario. Indeed, they ranged from 0.925 to 0.926 in the non-informative scenario, from 0.937 to 0.938 in the informative scenario, and from 0.971 to 0.972 in the optimistic scenario. The Bayes factors BF_{+-}, rounded to the nearest whole number, were consistent across methods and indicated moderate to strong evidence for H_+ over H_-: from 12.360 to 12.585 in the non-informative scenario, from 14.814 to 14.898 in the informative scenario, and from 8.501 to 8.616 in the optimistic scenario.[21] However, as noted by Dablander et al. (2022),[9] Bayes factors computed under the IBE approach are often optimistic.

Results of the LTT approach for the Move Ô case study

Figure 5 displays the simulated distributions of the parameters of interest, i.e. γ, ψ, π_1 and π_2, as well as the simulated distributions of the difference $\delta = \pi_2 - \pi_1$.

Table 3. Prior probabilities $P(H_+)$, posterior probabilities $P(H_+|data)$ and Bayes factors BF_{+-} estimated by the three implementations (i.e., our own simulation, bayesAB package, and MCMC with R2jags) of the IBE approach for the Move Ô case study in the three scenarios investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

		Non-informative	Informative	Optimistic
	Theoretical prior probability $P(H_+)$	0.500	0.500	
	Analytic posterior probability $P(H_+ \mid data)$	0.926	0.937	0.972
Simulation	Prior probability $P(H_+)$	0.500	0.503	0.800
	Posterior probability $P(H_+ \mid data)$	0.925	0.938	0.972
	Bayes factor BF_{+-}	12.360	14.859	8.616
bayesAB package	Posterior probability $P(H_+ \mid data)$	0.926	0.937	
	Bayes factor BF_{+-}	12.585	14.898	
MCMC with R2jags	Posterior probability $P(H_+ \mid data)$	0.926	0.937	0.971
	Bayes factor BF_{+-}	12.490	14.814	8.501

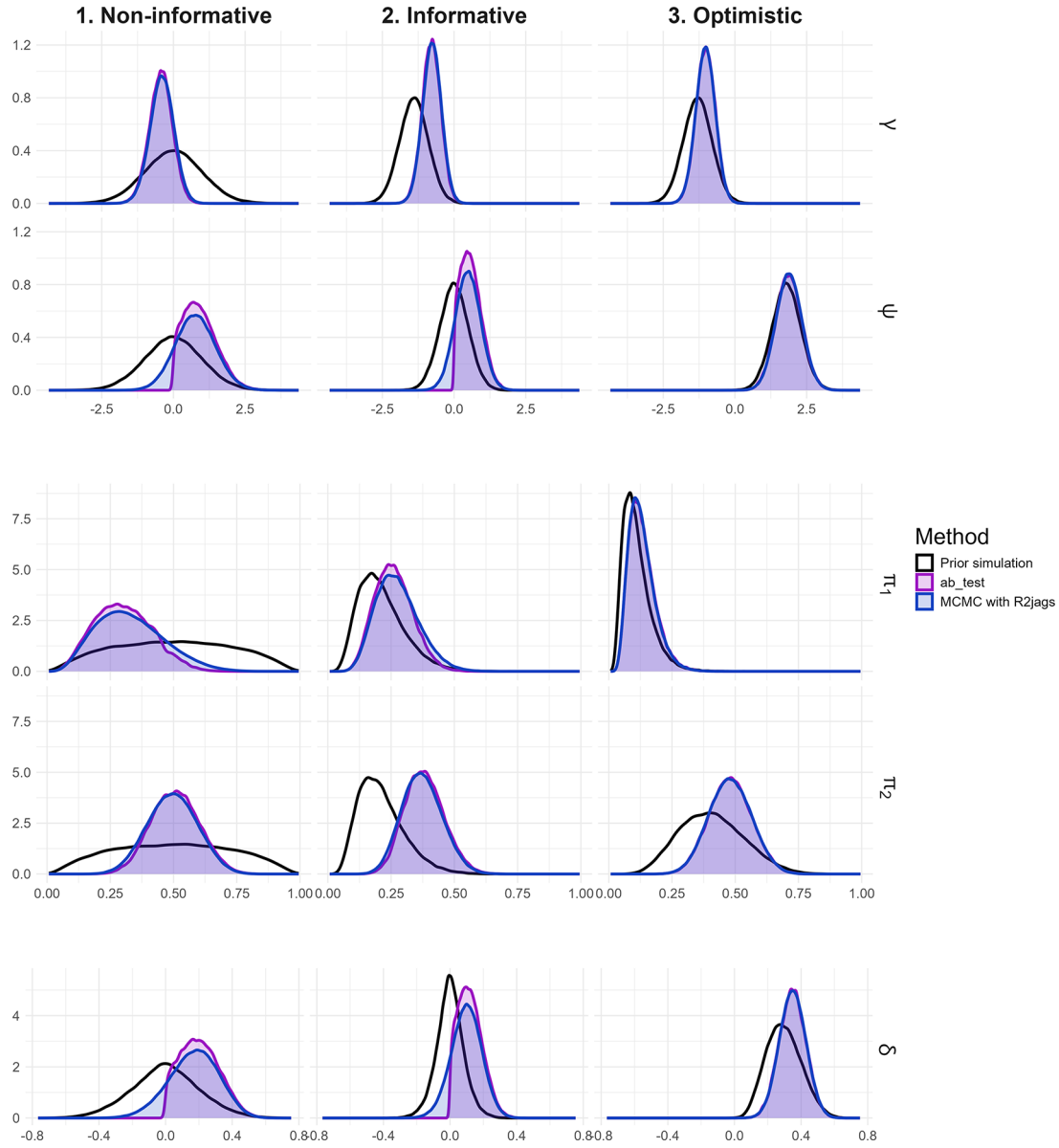


Figure 5. Distributions of the γ, ψ, π_1 and π_2 parameters and the difference $\delta = \pi_2 - \pi_1$ simulated under the prior and posterior settings for the Move Ô case study by the two implementations (ab_test from abtest package, and MCMC with R2jags) of the LTT approach for each scenario investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

The γ distributions depicted in Figure 5 shift towards negative values, with modes around -0.43 in the non-informative scenario, -0.77 in the informative scenario, and -1.00 in the optimistic scenario. This shift reflects an overall probability of success (i.e. the proportion of participants who were overweight or obese) ranging between 0.27 and 0.39, which was also a central value between π_1 and π_2. The main focus, however, is the ψ distributions, because they represent the magnitude of difference between π_1 and π_2. Their mode is positive in all scenarios, at approximately 0.75 in the non-informative scenario, 0.50 in the informative scenario, and 1.90 in the optimistic scenario. These values were the maximum likelihood for the logit difference according to the respective scenario.

The π_1 distributions show different modal values across the scenarios investigated: 0.28 in the non-informative scenario, 0.24 in the informative scenario, and 0.10 in the optimistic scenario. The greatest variability was observed in the non-informative scenario, while the smallest variability was observed in the optimistic scenario. The modal values of π_2 were

0.50 in the non-informative scenario, 0.37 in the informative scenario, and 0.48 in the optimistic scenario. In every scenario, the posterior distributions were more peaked than the corresponding prior distributions.

The modal values of δ shown in Figure 5 were 0.17 in the non-informative scenario, 0.10 in the informative scenario, and 0.34 in the optimistic scenario, thereby characterising the difference between π_1 and π_2. The highest distribution peak was observed in the informative scenario, whereas the lowest occurred in the non-informative scenario.

No major differences were observed between the two methods (i.e. *abtest* and MCMC with *R2jags*) except for the ψ and δ distributions in the non-informative and informative scenarios. These discrepancies arose because *abtest* uses a truncated distribution, whereas our MCMC implementation imposed no such constraint.

Table 4 reports the posterior probabilities and Bayes factors obtained from the two implementations of the LTT approach across the scenarios investigated. The hypothesis H_+ was compared first with H_0 and then with H_-. Table S2 complements these results by reporting the quantiles and 90% CI of the parameter distributions considered (γ, ψ, π_1, π_2, and $\delta = \pi_2 - \pi_1$).

For each scenario, both methods yielded comparable results for the main indicators, including posterior probabilities and Bayes factors. The posterior probabilities from the MCMC implementation with *R2jags* differed only slightly, at the third decimal place. Bayes factors, when rounded to the nearest whole number, were consistent across methods except in the optimistic scenario (Non-informative: $BF_{+0} \approx 2$ and $BF_{+-} \approx 6$; Informative: $BF_{+0} \approx 3$ and $BF_{+-} \approx 6$; Optimistic with *abtest*: $BF_{+0} \approx 11$ and $BF_{+-} \approx 14$; Optimistic with MCMC *R2jags*: $BF_{+0} \approx 9$ and $BF_{+-} \approx 10$), leading to the same substantive conclusions. More noticeable differences were observed for some quantiles, especially for ψ and δ (Table S2).

Table 4. Prior probabilities, posterior probabilities and Bayes factors estimated by the two implementations (abtest package, and MCMC with R2jags) of the LTT approach to compare H_+ over H_0 and H_+ over H_- for the Move Ô case study in the three scenarios investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

			Non-informative	Informative	Optimistic
H_+ over H_0	Prior probabilities	Prior probability $P(H_+)$	0.500	0.500	0.700
		Prior probability $P(H_0)$	0.500	0.500	0.300
	abtest package	Posterior probability $P(H_+ \mid data)$	0.689	0.733	0.961
		Posterior probability $P(H_0 \mid data)$	0.311	0.267	0.039
		Bayes factor BF_{+0}	2.214	2.748	10.621
	MCMC with R2jags	Posterior probability $P(H_+ \mid data)$	0.686	0.730	0.952
		Posterior probability $P(H_0 \mid data)$	0.314	0.270	0.048
		Bayes factor BF_{+0}	2.186	2.705	8.557
H_+ over H_-	Prior probabilities	Prior probability $P(H_+)$	0.500	0.500	0.700
		Prior probability $P(H_-)$	0.500	0.500	0.300
	abtest package	Posterior probability $P(H_+ \mid data)$	0.866	0.861	0.970
		Posterior probability $P(H_- \mid data)$	0.134	0.139	0.030
		Bayes factor BF_{+-}	6.482	6.170	13.876
	MCMC with R2jags	Posterior probability $P(H_+ \mid data)$	0.865	0.861	0.957
		Posterior probability $P(H_- \mid data)$	0.135	0.139	0.043
		Bayes factor BF_{+-}	6.434	6.170	9.548

Table 5. Robustness checking with the plot_robustness function from abtest package for the Move Ô case study: ranges of μ_ψ and σ_ψ inputs and Bayes factor BF_{+0} range.

	μ_ψ range input	σ_ψ range input	BF_{+0} range output
1. Non-informative	[0.0, 2.0]	[0.1, 2.0]	[1.14, 3.16]
2. Informative	[0.5, 2.5]	[0.1, 2.0]	[3.00, 15.24]
3. Optimistic	[0.5, 2.5]	[0.1, 2.0]	[2.89, 12.62]

As noted above, this can be attributed to the use of truncated normal distributions for ψ in *abtest*, whereas no constraint was applied in the MCMC *R2jags* implementation.

The *plot_robustness* function from the *abtest* package facilitates robustness assessment by producing a heatmap of Bayes factors across varying prior parameters for $\psi \sim \mathcal{N}\left(\mu_\psi, \sigma_\psi^2\right)$. Table 5 reports the minimum and maximum Bayes factor values obtained with this function. The Bayes factors varied only modestly and remained within the same order of magnitude, suggesting that the conclusions were reasonably robust. They indicated anecdotal to moderate evidence in the non-informative scenario (from 1.14 to 3.16 with the maximum observed for $\mu_\psi = 1.49$ and $\sigma_\psi = 0.10$), anecdotal to strong evidence in the informative scenario (from 3.00 to 15.24 with the maximum observed for $\mu_\psi = 2.00$ and $\sigma_\psi = 0.10$), and values of similar order of magnitude in favour of H_+ in the optimistic scenario (from 2.89 to 12.62 with the maximum observed for $\mu_\psi = 2.29$ and $\sigma_\psi = 0.10$).[20–22]

Results of the data alterations

Preliminary remarks on data alteration

We next examine how data alterations affect the results. The first alteration under consideration consists of adding one observation. This additional observation can be either a success or a failure in Group 1, producing an estimate $\widehat{\pi}_1$ that differs slightly from the original one, or a success or a failure in Group 2, similarly modifying $\widehat{\pi}_2$. Adding a single observation allows us to distinguish between alterations that are "favourable" to H_+, i.e. alterations that strengthen the evidence for H_+, and "unfavourable alterations", i.e. alterations that weaken the evidence for H_+. Adding a success (i.e. an additional overweight or obese adolescent) in Group 1 (where at least one parent is not overweight or obese) or a failure (i.e. an adolescent who is not overweight or obese) in Group 2 constitutes an "unfavourable alteration", whereas adding a failure in Group 1 or a success in Group 2 constitutes a "favourable alteration". The second alteration considered consists of doubling the sample size while preserving the originally observed proportions. In that case, the estimates $\widehat{\pi}_1$ and $\widehat{\pi}_2$ remain unchanged, but the evidence supporting the hypothesis favoured by the original data is expected to increase.

The results presented in the previous section showed that the different implementations of the IBE and LTT approaches yielded equivalent conclusions. Accordingly, this section reports only the simulation results for the IBE approach and the *abtest* results for the LTT approach.

It is also important to note that within the frequentist framework, conducting multiple tests as data accumulate is not acceptable because it introduces bias by increasing the risk of a Type I error.[8] Accordingly, when we discuss altered dataset in this section - particularly when referring the frequentist framework such as Fisher's exact test - the reader should assume that no prior testing was performed. Each altered dataset should instead be regarded as coming from a separate study conducted with the same methodology, differing only in the specific alteration considered.

Table 6. Effects of the data alterations on $\widehat{\pi}_1$, $\widehat{\pi}_2$, two-sided and one-sided p values of the Fisher's exact test for the Move Ô case study.

	$\widehat{\pi}_1$ (%)	$\widehat{\pi}_2$ (%)	Two-sided p	One-sided p
Another success in Group 1	29	52	0.396	0.258
Another failure in Group 1	14	52	0.184	0.091
Another success in Group 2	17	55	0.173	0.117
Another failure in Group 2	17	50	0.197	0.160
Doubling the sample size	17	52	0.046	0.028

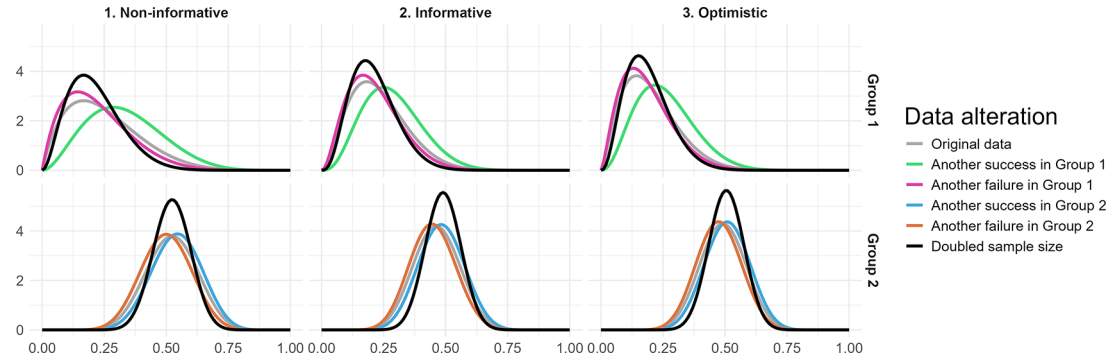


Figure 6. Posterior *βeta* distributions of the IBE approach applied to Groups 1 (upper row) and 2 (lower row) for the altered data from the Move Ô case study under each scenario investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

Influence of data alterations on the results of the Fisher's exact test

We begin the data-alteration analysis with Fisher's exact test. Table 6 shows the effects of the different data alterations on the estimates of π_1 ($\widehat{\pi}_1$) and π_2 ($\widehat{\pi}_2$), as well as on the two-sided and one-sided p-values.

Because Group 1 is small, even a minor change in the data can substantially affect $\widehat{\pi}_1$, the estimate of π_1. However, for every alteration involving the addition of only one observation, the p-values remained above 0.05, leading to the same conclusion: no significant difference between Groups 1 and 2. By contrast, doubling the sample size yielded $p < 0.050$, which led to rejection of H_0, i.e. a significant difference for the two-sided test and a significantly greater value of π_2 for the one-sided test.

Influence of data alterations on the results of the IBE approach

Under the IBE approach, Figure 6 shows how the estimated π_1 and π_2 distributions vary under the simulated data alterations. Adding a success in Group 1 shifted the density towards higher values, whereas adding a failure had little effect. This greater sensitivity reflects the small size of Group 1, which in the original dataset consisted predominantly of children who were not overweight or obese. In Group 2, by contrast, adding either a success or a failure produced only minor changes in the density. The larger sample size of Group 2 therefore made these alterations less influential than in Group 1. Doubling the sample size increased the height of the peaks and reduced the overlap between the distributions of Groups 1 and 2.

Table 7. Posterior probabilities $P(H_+|data)$ and Bayes factors BF_{+-} estimated by the implementation of the IBE approach with our own simulation on the altered data from the Move Ô case study in the three investigated scenarios: 1. Non-informative, 2. Informative, 3. Optimistic.

		Non-informative	Informative	Optimistic
	Theoretical prior probability $P(H_+)$	0.500	0.500	
Another success in Group 1	Posterior probability $P(H_+\|data)$	0.846	0.879	0.936
	Bayes factor BF_{+-}	5.467	7.176	3.667
Another failure in Group 1	Posterior probability $P(H_+\|data)$	0.952	0.954	0.981
	Bayes factor BF_{+-}	19.884	20.538	12.676
Another success in Group 2	Posterior probability $P(H_+\|data)$	0.938	0.950	0.977
	Bayes factor BF_{+-}	15.192	18.666	10.747
Another failure in Group 2	Posterior probability $P(H_+\|data)$	0.912	0.927	0.966
	Bayes factor BF_{+-}	10.353	12.519	7.042
Doubling the sample size	Posterior probability $P(H_+\|data)$	0.984	0.987	0.994
	Bayes factor BF_{+-}	61.195	76.202	42.734

Table 7 reports the effects of data alterations on posterior probabilities $P(H_+|data)$ and the Bayes factors BF_{+-}, while Table S10 provides additional information on indicators such as the quantiles of simulated π_1, π_2 and $\delta = \pi_2 - \pi_1$, together with 90% HDIs and 90% ETIs. Across all the scenarios investigated, evidence for H_+ over H_- weakened under unfavourable alterations and strengthened under favourable alterations. The strongest effect was observed when an observation was added to Group 1 because this group was smaller than Group 2. Bayes factors indicated moderate evidence (Non-informative: $BF_{+-} = 5.467$; Informative: $BF_{+-} = 7.176$; Optimistic: $BF_{+-} = 3.667$ when adding a success in Group 1) to strong evidence (Non-informative: $BF_{+-} = 19.884$; Informative: $BF_{+-} = 20.538$; Optimistic: $BF_{+-} = 12.676$ when adding a failure in Group 1). Doubling the sample size strengthened the indicators in favour of H_+, with higher posterior probabilities (Non-informative: $P(H_+|data) = 0.984$; Informative: $P(H_+|data) = 0.987$; Optimistic: $P(H_+|data) = 0.994$), larger Bayes factors (Non-informative: $BF_{+-} = 61.195$; Informative: $BF_{+-} = 76.202$; Optimistic: $BF_{+-} = 42.734$), and CIs for δ that excluded the null value.

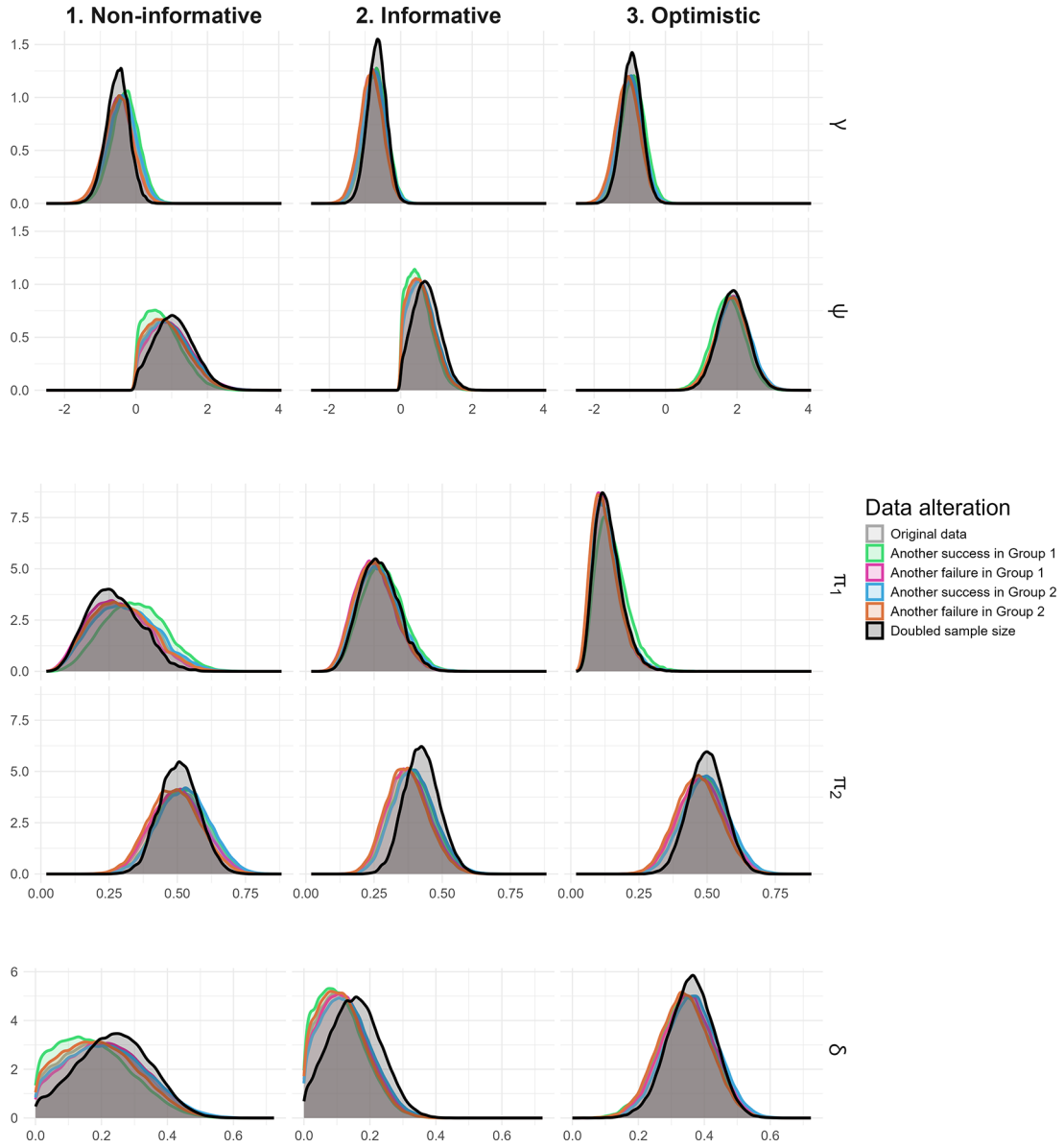


Figure 7. Posterior distributions of the γ, ψ, π_1 and π_2 parameters and the difference $\delta = \pi_2 - \pi_1$ simulated for the altered data from the Move Ô case study by the abtest implementation of the LTT approach for each scenario investigated: 1. Non-informative, 2. Informative, 3. Optimistic.

Table 8. Prior probabilities, posterior probabilities and Bayes factors estimated by the implementation of the LTT approach with the abtest package to compare H_+ over H_0 and H_+ over H_- on the altered data from the Move Ô case study in the three investigated scenarios: 1. Non-informative, 2. Informative, 3. Optimistic.

			Non-informative	Informative	Optimistic
Prior probabilities	H_+ over H_0	Prior probability $P(H_+)$	0.500	0.500	0.700
		Prior probability $P(H_0)$	0.500	0.500	0.300
	H_+ over H_-	Prior probability $P(H_+)$	0.500	0.500	0.700
		Prior probability $P(H_-)$	0.500	0.500	0.300
Another success in Group 1	H_+ over H_0	Posterior probability $P(H_+\|data)$	0.596	0.678	0.904
		Posterior probability $P(H_0\|data)$	0.404	0.322	0.096
		Bayes factor BF_{+0}	1.476	2.104	4.019
	H_+ over H_-	Posterior probability $P(H_+\|data)$	0.791	0.811	0.922
		Posterior probability $P(H_-\|data)$	0.209	0.189	0.078
		Bayes factor BF_{+-}	3.787	4.302	5.038
Another failure in Group 1	H_+ over H_0	Posterior probability $P(H_+\|data)$	0.740	0.755	0.971
		Posterior probability $P(H_0\|data)$	0.260	0.245	0.029
		Bayes factor BF_{+0}	2.843	3.075	14.154
	H_+ over H_-	Posterior probability $P(H_+\|data)$	0.903	0.879	0.978
		Posterior probability $P(H_-\|data)$	0.097	0.121	0.022
		Bayes factor BF_{+-}	9.321	7.276	18.884
Another success in Group 2	H_+ over H_0	Posterior probability $P(H_+\|data)$	0.714	0.756	0.972
		Posterior probability $P(H_0\|data)$	0.286	0.244	0.028
		Bayes factor BF_{+0}	2.497	3.094	14.718
	H_+ over H_-	Posterior probability $P(H_+\|data)$	0.884	0.880	0.979
		Posterior probability $P(H_-\|data)$	0.116	0.120	0.021
		Bayes factor BF_{+-}	7.632	7.303	19.600
Another failure in Group 2	H_+ over H_0	Posterior probability $P(H_+\|data)$	0.667	0.720	0.952
		Posterior probability $P(H_0\|data)$	0.333	0.280	0.048
		Bayes factor BF_{+0}	2.002	2.569	8.459
	H_+ over H_-	Posterior probability $P(H_+\|data)$	0.849	0.849	0.962
		Posterior probability $P(H_-\|data)$	0.151	0.151	0.038
		Bayes factor BF_{+-}	5.613	5.631	10.921

Table 8. *Continued*

			Non-informative	Informative	Optimistic
Doubling the sample size	H_+ over H_0	Posterior probability $P(H_+ \mid data)$	0.855	0.872	0.991
		Posterior probability $P(H_0 \mid data)$	0.145	0.128	0.009
		Bayes factor BF_{+0}	5.887	6.817	48.706
	H_+ over H_-	Posterior probability $P(H_+ \mid data)$	0.967	0.958	0.994
		Posterior probability $P(H_- \mid data)$	0.033	0.042	0.006
		Bayes factor BF_{+-}	29.142	22.685	75.390

Influence of data alterations on the results of the LTT approach

Under the LTT approach, Figure 7 shows that the data alterations affected the simulated distributions of γ and ψ. The largest changes occurred in the ψ distribution when an additional success was added to Group 1, in both non-informative and informative scenarios. This produced a leftward shift of the distribution, and the use of a truncated positive distribution led to a higher peak. Doubling the sample size also increased the peaks for the parameter distributions.

Although the overall shapes of the π_1 and π_2 distributions remained largely unchanged, the most noticeable alterations were observed for π_2 when the sample size was doubled (Figure 7). These changes were consistent with expectations. Unfavourable alterations shifted the π_1 distribution to the right and the π_2 distribution to the left. Conversely, favourable alterations shifted the π_1 distribution to the left and the π_2 distribution to the right. Doubling the sample size shifted the π_2 distribution to the right and increased its peak.

As noted above, unfavourable alterations shifted the δ distribution to the left, thereby weakening evidence in support of the $H_+ : \pi_1 < \pi_2$ hypothesis. By contrast, favourable alterations shifted it to the right, thereby strengthening the evidence for the H_+ hypothesis (Figure 7). Since the original data supported H_+, doubling the sample size reinforced this result through a rightward shift and a higher peak.

Table 8 reports the updated posterior probabilities and Bayes factors after the data alterations, together with the hypotheses compared. Table S11 provides supplementary information on the quantiles and CIs of the parameter distributions. The findings were consistent with those obtained under the IBE approach. In all cases, unfavourable alterations (adding a success in Group 1 and a failure in Group 2) reduced the evidence for H_+ over H_0 or H_-, with Bayes factors indicating weak to moderate evidence for H_+ over H_0 or H_- (Non-informative: 1.476 to 5.613; Informative: 2.104 to 5.631; Optimistic: 4.019 to 10.921). Favourable alterations (adding a failure in Group 1 and a success in Group 2) strengthened the evidence in favour of H_+, with Bayes factors indicating weak to strong evidence for H_+ over H_0 or H_- (Non-informative: 2.497 to 9.321; Informative: 3.075 to 7.303; Optimistic: 14.154 to 19.600). When the sample size was doubled, Bayes factors indicated moderate to very strong evidence for H_+ over H_0 or H_- (Non-informative: 5.887 and 29.142; Informative: 6.817 and 22.685; Optimistic: 48.706 and 75.390).

Discussion and conclusion

In this study, we applied both the IBE and LTT approaches to a case study involving an understudied population living in a PICT. Although both approaches were implemented, the literature recommends giving preference to the LTT approach.[9] We reviewed, illustrated, and discussed several related indicators in order to assess the possible association between the weight status of children and that of their parents. Even if the present study was not intended as an apologia for the Bayesian framework, it shows that these approaches provide valuable information that deepens the conclusion obtainable from a frequentist test such as Fisher's exact test, especially when applied to small datasets.[40] Rather than ending with a closed conclusion such as "non-significant" or "H_1 rejected", the analysis makes it possible to distinguish between "evidence of absence", "absence of evidence", and even stronger forms of support.[8,10] We discuss these points below.

The supplementary materials (Tables S3 to S9 and S12 to S16; Figures S1 to S5) extend the present results with similar findings from another case study conducted on Lifou Island in New Caledonia, another PICT. Although the two case studies addressed similar substantive topic, the two datasets were complementary because of important differences

between them. Indeed, the sample size from the Move Ô study was smaller than in the Lifou study (Move Ô: $n = 27$; Lifou: $n = 48$), but the estimated difference $\widehat{\delta} = \widehat{\pi}_2 - \widehat{\pi}_1$ was larger in the first case study (Move Ô: $\widehat{\delta} = 0.35$; Lifou: $\widehat{\delta} = 0.23$), as was the effect size h (Move Ô: $h = 0.78$; Lifou: $h = 0.59$). This explains why the highlighted evidence was weaker in the supplementary case study than in the present one, and why the results were less affected by the data alterations.

This study made it possible to compare the meaning and outputs of the frequentist Fisher's exact test with those of the Bayesian A/B testing approaches (IBE and LTT). While the frequentist framework focuses on comparing the observed data with the distribution expected under H_0, the Bayesian framework instead updates both H_0 and the alternative hypothesis, which may be H_1, H_-, or H_+. For the primary dataset considered here, Fisher's exact test yielded an absence of evidence favouring H_0 over H_1 or H_+, whereas the Bayes factors calculated from the IBE approach indicated moderate to strong evidence for H_+ over H_- and the Bayes factors from the LTT approach indicated negligible (i.e. essentially absent) to moderate evidence for H_+ over H_0.[8,9] Bayes factors are known to be sensitive to the choice of prior, especially in the small-sample datasets.[8,31,40,47] However, we selected three reasonable prior specifications in relation to the present context, and unlike in some previous simulation studies, the substantive conclusions did not change markedly because the Bayes factors never favoured H_0.[47]

Other indicators, such as the δ distribution and the associated 90% CI, as well as the shift from prior to posterior distributions and the corresponding posterior probabilities, provided additional insight. In particular, even when the CI for δ included the null value (i.e. in the non-informative and informative scenarios), most of the mass of the δ distribution remained above 0, yielding high posterior probabilities in favour of H_+. These probabilities exceeded 0.90 for the IBE approach; in the LTT approach they ranged from 0.65 to 0.96 when comparing H_0 and H_+, and exceeded 0.82 when comparing H_- and H_+ (Tables 3, 4, S5, and S6). Uncertainty remains, and additional data would strengthen the evidence, but the convergence of these indicators suggests that concluding that H_+ is highly probable is not unreasonable.[31]

Regarding the data alterations, doubling the sample size was intended to emphasize the effect of sample size on the results and conclusions while leaving the effect size unchanged. The p-values from Fisher's exact test fell below the conventional rejection threshold ($p < 0.050$), and the Bayesian indicators suggested stronger evidence in favour of H_+ over the competing hypotheses. Importantly, the overall shapes of the posterior distributions changed little, except that they became narrower and more peaked. The Bayes-factor evidence for H_+ strengthened to strong to very strong under the IBE approach (Tables 7 and S15). Under the LTT approach, evidence for H_+ over H_0 was moderate to very strong, whereas evidence for H_+ over H_- was strong to very strong (Tables 8 and S16). This type of alteration also illustrates that, when the effect size is larger, increasing the amount of data consistent with the observed effect - in this case, data in favour of H_+ - strengthens the evidence more readily. For example, when the LTT approach was used to compare H_+ and H_0, Bayes factors for the original data ranged from 2 to 11 in the Move Ô case study (with strong effect size) and from 2 to 3 in the Lifou case study (with moderate effect size) (Tables 4 and S6); after doubling the sample size, the corresponding ranges were from 6 to 49 and from 5 to 11, respectively (Tables 8 and S14).

By contrast, adding a single observation led to slight changes in both sample size and effect size. As in the Results section, we distinguish favourable alterations, which provide additional support for H_+, from unfavourable alterations, which do not. The conclusions based on Fisher's exact test remained unchanged regardless of the alteration considered ($p \geq 0.050$). The main Bayes-factor conclusions also remained unchanged, although the strength of the evidence for H_+ over competing hypotheses could be slightly increased or decreased depending on the alteration because of the interpretive thresholds used.[20,21] It is also noteworthy that, if one estimates the proportion of favourable alterations from the available data, the behaviour of the sampling process should differ between the two case studies. Although this lies beyond the scope of the present paper, stronger evidence can be anticipated in the Move Ô case study ($16/27 \approx 59\%$ favourable alterations), whereas it would be weaker in the supplementary Lifou case study ($22/48 \approx 46\%$ favourable alterations).

It is important to acknowledge that researchers who are less familiar with the Bayesian framework may find potentially divergent conclusions based on posterior probabilities and Bayes factors frustrating. Etz & Vandekerckhove (2018) clarified the complementary roles of these two indicators.[19] Posterior probabilities are useful "to determine our total belief after taking into account the data and to draw conclusions", whereas Bayes factor serves as "a learning factor that tells us how much evidence the data have delivered".[19] In the case study examined here, no counterintuitive result was observed. The data, along with the resultant high posterior probabilities, were consistent with existing theories and previous case studies.[42–44] However, the data provided only limited additional evidence - ranging from negligible to moderate - under the LTT approach, for example.

This paper showed that the application of statistical tools from the Bayesian framework can offer insights that complement those obtained under the conventional frequentist framework. Our aim is not to replace existing practices, but to advocate the inclusion of this additional information in order to obtain a fuller understanding of the situations under study.[8] We believe that this approach is particularly valuable for research on understudied populations that are challenging to access. Indeed, interventions in such populations may yield modest yet clinically and developmentally meaningful changes in anthropometric or behavioural outcomes, but these effects are often difficult to demonstrate within the "traditional frequentist framework".

Consent

We obtained informed written consent from all parents for their participation and their children's participation before their children entered the study. The research met the legal requirements and the Declaration of Helsinki, and the protocols[18,49] were approved by the Consultative Ethics Committee of New Caledonia: CCE 2018–06 01.

Disclosure

Deepl and Paperpal AI assistant were used to improve readability and language of the work and not used to replace researcher tasks such as producing scientific insights, analysing and interpreting data, or drawing scientific conclusions. Human oversight and control have been applied to the text editing improvement, and authors have carefully reviewed and edited the result because AI assistants can generate authoritative-sounding output that can be incorrect, incomplete, or biased.

Supplementary materials

The supplementary materials are included in deposit containing the underlying data.

Data and software availability

Underlying data

All supplementary materials as well as data and R code necessary to reproduce this manuscript can be found at https://doi.org/10.5281/zenodo.17982872.[48] The content of this deposit is available under the terms of the CC-BY 4.0 licence.

Extended data

No extended data.

Acknowledgments

We would like to express our sincere gratitude to Maïwenn Moreau from the Pacific Community (SPC), as well as the staff of the *Direction de l'Enseignement Catholique* (Catholic Education Department), the regional committee of the *Union Générale Sportive de l'Enseignement Libre* (General Sports Union for Independent Education), and the *Agence de Santé de Wallis-et-Futuna* (Wallis and Futuna Health Agency) for their participation in data collection and their invaluable logistical support regarding the Move Ô case study. We also thank Paul Zongo, the *Autorités coutumières des Îles Loyauté* (Loyalties Islands Customary Authorities) on Lifou Island and the families involved in the supplementary Lifou case study.

References

1. Galy O, Frayon S, Goldin M, *et al.*: **Generational issues in linking family farming production, traditional food in diet, physical activity and obesity in Pacific Islands countries and territories: the case of the Melanesian population on Lifou Island.** *Open Res Europe.* 2022; **1**: 135.
Publisher Full Text

2. Gwynn JD, Flood VM, D'Este CA, *et al.*: **The reliability and validity of a short FFQ among Australian Aboriginal and Torres Strait Islander and non-Indigenous rural children.** *Public Health Nutrition.* 2011; **14**: 388–401.
PubMed Abstract | **Publisher Full Text**

3. Hoffmann T, Hofman A, Wagenmakers E-J: **Bayesian tests of two proportions: A tutorial with R and JASP.** *Methodology.* 2022; **18**: 239–277.
Publisher Full Text

4. Kessaram T, McKenzie J, Girin N, *et al.*: **Overweight, obesity, physical activity and sugar-sweetened beverage consumption in adolescents of Pacific islands: results from the Global School-Based Student Health Survey and the Youth Risk Behavior Surveillance System.** *BMC Obes.* 2015; **2**: 34.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Wasserstein RL, Lazar NA: **The ASA Statement on *p* -Values: Context, Process, and Purpose.** *The American Statistician.* 2016; **70**: 129–133.
Publisher Full Text

6. Cohen J: *Statistical Power Analysis for the Behavioral Sciences.* Routledge;0th ed 2013.
Publisher Full Text

7. Sullivan GM, Feinn R: **Using Effect Size—or Why the P Value Is Not Enough, Journal of Graduate.** *Medical Education.* 2012; **4**: 279–282.
Publisher Full Text

8. Keysers C, Gazzola V, Wagenmakers E-J: **Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence.** *Nat Neurosci.* 2020; **23**: 788–799.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Dablander F, Huth K, Gronau QF, *et al.*: **A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions.** *Statistics in Medicine.* 2022; **41**: 1319–1333.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Altman DG, Bland JM: **Statistics notes: Absence of evidence is not evidence of absence.** *BMJ.* 1995; **311**: 485.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Berndt AE: **Sampling Methods.** *J Hum Lact.* 2020; **36**: 224–226.
Publisher Full Text

12. Hasan Polas MR: **Sample selection in social science research: A holistic approach to methodological rigor.** *Sustain. Econ.* 2024; **2**: 31.
Publisher Full Text

13. **Cour des comptes, Le système scolaire en Polynésie française et en Nouvelle-Calédonie: un effort de l'État important, une efficience à améliorer.** *Rapport public annuel 2016, Cour des comptes.* 2016; p. 41. (accessed March 12, 2024).
Reference Source

14. Friesen W, Bedford R, Underhill-Sem Y: *Country Reports on population dynamics in the Pacific, 1950–2050, Bedford Consulting&Waipapa Taumata Rau.* University of Auckland; 2023.
Reference Source

15. Lal S, Singh R, Chand R, *et al.*: **Projecting populations for major Pacific Island countries with and without COVID-19: pro-active insights for population policy.** *J Pop Research.* 2022; **39**: 257–277.
PubMed Abstract | Publisher Full Text | Free Full Text

16. (NCD-RisC) NCD Risk Factor Collaboration: **Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants.** *The Lancet.***387**(2016): 1377–1396.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Bertrand-Protat S, Chen J, Jonquoy A, *et al.*: **Prevalence, causes and contexts of childhood overweight and obesity in the Pacific region: a scoping review.** *Open Res Europe.* 2023; **3**: 52.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Bertrand-Protat S, Wattelez G, Frayon S, *et al.*: **Encourager la pratique de l'activité physique chez les enfants du Pacifique, un moyen efficace pour prévenir l'obésité infantile ?.** *Étude pilote en milieu scolaire sur le territoire des îles de Wallis-et-Futuna:, Santé Publique.* 2024; **36**: 71–81.
Publisher Full Text

19. Etz A, Vandekerckhove J: **Introduction to Bayesian Inference for Psychology.** *Psychon Bull Rev.* 2018; **25**: 5–34.
PubMed Abstract | Publisher Full Text

20. Stefan AM, Gronau QF, Schönbrodt FD, *et al.*: **A tutorial on Bayes Factor Design Analysis using an informed prior.** *Behav Res.* 2019; **51**: 1042–1058.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Kass RE, Raftery AE: **Bayes Factors.** *Journal of the American Statistical Association.* 1995; **90**: 773–795.
Publisher Full Text

22. Jeffreys H: *The theory of probability.* Oxford: Oxford University Press; 2023.
Publisher Full Text

23. WHO Global InfoBase team: *Surveillance of chronic disease risk factors: country-level data and comparable estimates.* Geneva: WHO; 2005.

24. Cole TJ, Lobstein T: **Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity.** *Pediatric Obesity.* 2012; **7**: 284–294.
PubMed Abstract | Publisher Full Text

25. Sprent P: **Fisher Exact Test.** Lovric M, editor. *International Encyclopedia of Statistical Science.* Berlin Heidelberg, Berlin, Heidelberg: Springer; 2011; pp. 524–525.
Publisher Full Text

26. Lee MD, Vanpaemel W: **Determining informative priors for cognitive models.** *Psychon Bull Rev.* 2018; **25**: 114–127.
Publisher Full Text

27. Etz A, Wagenmakers E-J: **J. B. S. Haldane's Contribution to the Bayes Factor Hypothesis Test.** *Statist. Sci.* 2017; **32**.
Publisher Full Text

28. Van Doorn J, Matzke D, Wagenmakers E-J: **An In-Class Demonstration of Bayesian Inference.** *Psychology Learning & Teaching.* 2020; **19**: 36–45.
Publisher Full Text

29. Gronau QF, Akash Raj KN, Wagenmakers E-J: **Wagenmakers, Informed Bayesian Inference for the A/B Test.** *J. Stat. Soft.* 2021; **100**.
Publisher Full Text

30. R Core Team: **A language and environment for statistical computing.** 2024.
Reference Source

31. van Ravenzwaaij D, Etz A: **Corrigendum: Simulation Studies as a Tool to Understand Bayes Factors.** *Advances in Methods and Practices in Psychological Science.* 2021; **4**: 25152459211061266.
Publisher Full Text

32. Portman F: **bayesAB: Fast Bayesian Methods for AB Testing.** 2021. (accessed January 16, 2026).
Reference Source

33. Su Y-S, Yajima M: **R2jags: Using R to Run "JAGS,".** 2024. (accessed January 16, 2026).
Reference Source

34. Van Ravenzwaaij D, Cassey P, Brown SD: **A simple introduction to Markov Chain Monte–Carlo sampling.** *Psychon Bull Rev.* 2018; **25**: 143–154.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Del Moral P, Doucet A, Jasra A: **Sequential Monte Carlo Samplers, Journal of the Royal Statistical Society Series B.** *Statistical Methodology.* 2006; **68**: 411–436.
Publisher Full Text

36. Geweke J: **Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.** Bernardo JM, Berger JO, Dawid P, *et al.*, editors. *Bayesian Statistics 4.* Oxford University PressOxford; 1992; pp. 169–193.
Publisher Full Text

37. Makowski D, Ben-Shachar M, D.: **Lüdecke, bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework.** *JOSS.* 2019; **4**: 1541.
Publisher Full Text

38. Kass RE, Vaidyanathan SK: **Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions.** *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 1992; **54**: 129–144.
Publisher Full Text

39. Hailpern SM, Visintainer PF: **Odds Ratios and Logistic Regression: Further Examples of their use and Interpretation.** *The Stata Journal: Promoting Communications on Statistics and Stata.* 2003; **3**: 213–225.
Publisher Full Text

40. Vanpaemel W: **Prior sensitivity in theory testing: An apologia for the Bayes factor.** *Journal of Mathematical Psychology.* 2010; **54**: 491–498.
Publisher Full Text

41. Frayon S, Wattelez G, Paufique E, *et al.*: **Overweight in the pluri-ethnic adolescent population of New Caledonia: Dietary patterns, sleep duration and screen time.** *The Lancet Regional Health - Western Pacific.* 2020; **2**: 100025.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Bahreynian M, Qorbani M, Khaniabadi BM, *et al.*: **Association between Obesity and Parental Weight Status in Children and Adolescents.** *Jcrpe.* 2017; **9**: 111–117.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Crossman A, Anne Sullivan D, Benin M: **The family environment and American adolescents' risk of obesity as young adults.** *Social Science & Medicine.* 2006; **63**: 2255–2267.
PubMed Abstract | Publisher Full Text

44. Maia C, Braz D, Fernandes HM, *et al.*: **The Impact of Parental Behaviors on Children's Lifestyle, Dietary Habits, Screen Time, Sleep Patterns, Mental Health, and BMI: A Scoping Review.** *Children.* 2025; **12**: 203.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Wagenmakers E-J, Lodewyckx T, Kuriyal H, *et al.*: **Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method.** *Cognitive Psychology.* 2010; **60**: 158–189.
Publisher Full Text

46. Heck DW: **A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters.** *Brit J Math & Statis.* 2019; **72**: 316–333.
Publisher Full Text

47. Liu CC, Aitkin M: **Bayes factors: Prior sensitivity and model generalizability.** *Journal of Mathematical Psychology.* 2008; **52**: 362–375.
Publisher Full Text

48. Wattelez G: **Supplementary materials, data and code of "Using Bayesian test relationship between categorical random variables on small sample sizes: application on the relationship between the weight status of parents and their children in understudied populations living in Pacific Island Countries and Territories,".** 2025.
Publisher Full Text

49. Galy O, Frayon S, Goldin M, *et al.*: **Protocol of the pilot study: generational issues in linking family farming production, traditional food in diet, physical activity and obesity in Pacific Islands countries and territories: the case of the Melanesian population on Lifou Island v1.** 2021.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 26 September 2026

https://doi.org/10.21956/openreseurope.26035.r79851

© **2026 Johnson V.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Valen E. Johnson
Texas A&M University, College Station, USA

This article provides a case study for the application of Bayesian testing and inference procedures for categorical variables. Specifically, it describes statistical methodology for testing potential differences between binomial proportions.
The case study data were collected from a Move Ô protocol conducted on Wallis Island that measured BMI values for 52 Grade 5 students. Parents of these students were asked to complete a survey of father and mother's height and weight, from which their BMI's were calculated. Twenty-seven participants had complete height and weight data for both parents and child and were subsequently included in the analyses. Children with two overweight or obese parents (BMI>25) comprised Group 2, and the remaining children were placed in Group 1. Statistical analyses focused on assessing differences between $\pi 1 \pi_1$ and $\pi 2 \pi_2$, the population probabilities that children in the two groups were overweight or obese. The observed proportions were $\pi 1=1/6$ $\hat{\pi_1} = 1/6$ and $\pi 2=11/21 \hat{\pi_2} = 11/21$.

The two-sided Fisher's test for these data yielded a p-value of 0.182, which was deemed not significant. To provide more informative summaries of evidence for the effect of having overweight or obese children, the authors investigated four hypotheses from a Bayesian perspective.

The four hypotheses, stated in normal frequentist format, were $H0:\pi 1=\pi 2, H_0: \pi_1 = \pi_2$, $H1:\pi 1\neq\pi 2, H_1: \pi_1 \neq \pi_2$, $H-:\pi 1>\pi 2, H_- : \pi_1 > \pi_2$, and $H+:\pi 1<\pi 2. H_+ : \pi_1 < \pi_2$. Of course, these specifications are not complete from a Bayesian perspective because Bayesian tests require the specification of a complete probability model for the data under each hypothesis. To complete these specifications, the authors use formulations provided in Doblander et al. 2022. Specifically, they propose independent binomial models for data, and either (i) independent beta priors for independent beta estimation (IBE), or (ii) normal priors on parameters in a logistic transformation of the binomial success probabilities in the logistic transformation testing (LTT) model. However, the specification of the IBE model is not consistent with the standard Bayesian formulation and

does not appear to be consistent with the numerical values reported in Table 3, the prior density plots in Figure 3, or the summaries provided in Table 2.

To understand the problem with the IBE results, note that the IBE model assumes that πi~beta$\alpha i,\beta i$ $\pi_i \sim beta(\alpha_i,\beta_i)$ for $i=1,2$ $i = 1,2$, ***independently*** for all four hypotheses (with the implicit assumption that $\alpha i,\beta i>0$). $\alpha_i,\beta_i > 0)$. However, this specification for the success probabilities is not consistent with either $H-$ H_- or $H+$ H_+, since in both cases the prior distributions assign positive probability to the both events $\pi 1>\pi 2$ $\pi_1 > \pi_2$ and $\pi 1<\pi 2$ $\pi_1 < \pi_2$. Although it is not generally a requirement that priors used to define Bayes factors be mutually singular, mutual singularity is typically assumed for directional hypotheses (i.e., $\theta>0$ *versus* $\theta<0$, $\theta > 0 \ versus \ \theta < 0,$ or $\theta 1<\theta 2$ $\theta_1 < \theta_2$). Thus, the IBE formulation with independent prior beta distributions is not appropriate for testing these hypotheses. The numerical values of the Bayes factors reported in Table 3 instead seem to have been computed under priors that were specified as

$H-:prior\pi 1,\pi 2\propto$ beta$\pi 1;\alpha 1,\beta 1 * beta\pi 2;\alpha 2,\beta 2 * Ind\pi 1>\pi 2,$
$H_-: prior(\pi_1,\pi_2) \propto beta(\pi_1;\alpha_1,\beta_1) * beta(\pi_2;\alpha_2,\beta_2) * Ind(\pi_1 > \pi_2),$

and
$H+:prior\pi 1,\pi 2\propto$ beta$\pi 1;\alpha 1,\beta 1*beta\pi 2;\alpha 2,\beta 2*Ind\pi 1<\pi 2$
$H_+: prior(\pi_1,\pi_2) \propto beta(\pi_1;\alpha_1,\beta_1) * beta(\pi_2;\alpha_2,\beta_2) * Ind(\pi_1 < \pi_2)$

Here, $beta\pi;\alpha,\beta$ $beta(\pi;\alpha,\beta)$ denotes a beta density with parameters α,β (α,β) and $Indx>y$ $Ind(x > y)$ is the indicator function of whether $x>y$ $x > y$ or not.

Note that specifying a single prior for $\theta 1$ and $\theta 2$ θ_1 and θ_2 defines only one model [e.g., $prior\pi 1,\pi 2 \propto beta\pi 1;\alpha 1,\beta 1*beta\pi 2;\alpha 2,\beta 2$ $prior(\pi_1,\pi_2) \propto beta(\pi_1;\alpha_1,\beta_1) * beta(\pi_2;\alpha_2,\beta_2)$], and a second model is necessary for calculating a Bayes factor.
A similar confusion appears on page 8, where the authors state that "this [IBE] test does not allow the experimenter to assess the exact equality of proportions." This statement may underlie the authors' focus on testing $H+$ H_+ versus $H-$ H_- and is true if independent beta priors are assumed to define $H0$. H_0. However, the implicit modification of the IBE framework to test $H0$ H_0 is to assume that, say, $\pi 1$~beta$\alpha 1,\beta 1$ $\pi_1 \sim beta(\alpha_1,\beta_1)$, and that $\pi 2=\pi 1$. $\pi_2 = \pi_1.$ This is the assumption made

in Dablander et al. 2022, where Bayes factors between $H0$ H_0 and $H1$ H_1 are studied extensively. For these data, if a uniform prior is assumed for $\pi 1=\pi 2$, $\pi_1 = \pi_2,$ then $BF10\approx 1.49$, $BF_{10} \approx 1.49,$

which is not unexpected given the modest p-value based on Fisher's test. Similarly, *BF+0*≈2.76 $BF_{+0} \approx 2.76$ if a truncated uniform prior is assumed under *H+*. H_+.
In my opinion, the Bayes factors reported from the IBE method are inflated, and the cause of this inflation is discussed thoroughly by Dablander et al. (2022).
The discussion of the effects of informative priors is interesting, but it should be noted that the informative priors imposed are nearly as informative as the data themselves. For example, the informative prior for the IBE model has an effective sample size of 7 (recall that *n1=6* $n_1 = 6$), and for the optimistic prior on Group 1 is 5.5. Quantifying the effective sample size of the priors in the LTT model is more difficult, but these also seem quite informative. If these priors are indeed supported by previous studies, then it seems that a hierarchical model should be specified and used for testing. Similar comments apply to "doubling the data."
I did not find the discussion of altering the data particularly useful. Randomness in the data is already cooked into the sampling density. Sensitivity to this assumption might be better assessed by considering other sampling models, e.g., the hypergeometric likelihood function used to construct Fisher's test. And, of course, the most serious issue concerning the data is likely the non-ignorable nonresponse, which, for example, may have made obese parents less likely to participate in the study if their children were obese.
In summary, I welcome the presentation of Bayesian analyses and hypothesis tests based on Bayes factors, and I agree that Bayes factors, accompanied by a clear description of the sampling models and priors upon which they are based, provide more information than the simple report of p-values. Of course, the value of the resulting Bayesian tests depends on the appropriateness of their specification, and in particular the coherence between the statistical hypotheses and substantive questions asked and the probability models used to answer those questions.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
No

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian statistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 01 September 2026

https://doi.org/10.21956/openreseurope.26035.r79849

© 2026 Malard-Adam J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Julien Jean Malard-Adam
Universite de Montpellier, Montpellier, Occitanie, France

This is a very interesting study that analyses the impact of adopting a Bayesian statistic framework in the context of small observation sizes. The need for the study and its impacts are well presented and justified. My recommendations deal mostly with clarifications to the methodology and analysis sections of the paper.

Abstract:
The background states that "rejecting the null hypothesis does not provide evidence of the absence of an effect." I believe that the authors meant to say "not rejecting the null hypothesis."
What is the "probability of success"?

Introduction:
"the Bayesian approach also requires an explicit formulation of the alternative hypothesis" -> I would rather say, "...of prior knowledge." Bayesian approaches do not require formulation of null or alternative hypotheses *per se*.
"distinguish between absence of evidence and evidence of absence" -> I very much like this formulation!

Methods:
For the optimistic scenario, selecting priors based on data from elsewhere in the world would make for a more justifiable selection of prior distributions.
Eq. 12: There seems to be something missing before the multiplication sign on the right side of the equation.
"yielded $p \approx 0.139$; again $H0$ was not rejected" -> there should be a comma after "again"
It is not clear to me exactly what the "own simulation" analysis refers to. This should be clearly indicated - how were these results obtained, and from which equation(s) and input data? How do these compare, theoretically and practically, to the methods in the bayesAB and MCMC approaches? It is also not immediately clear to me how the bayesAB and MCMC approaches differ.
The codes provided do not seem to include the code used to generate the figures presented in the article.

Analysis:
Figure 5: Delta has a very steep drop-off at 0 in the ab_test method graphs. Is this an artefact or error? The text states that this is due to a truncated distribution; does this mean the the priors in this model preclude a negative value for delta?
It is not clear to me how the MCMC models comparing H0 to H+ and the MCMC models comparing H+ to H- were formulated differently. In an MCMC approach, both models (H+ and H0 or H-) would be fitted separately (and then possibly compared); however, in Table 4, posterior probabilities of P(H+ | data) of the same H+ model differ between the two scenarios.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian inference (MCMC). Review by a reviewer with expertise in the other bayesian approaches used here (LTT, IBE) would be welcome.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.