HHS Public Access
Author manuscript
Stat Med. Author manuscript; available in PMC 2023 March 15.

Published in final edited form as:
Stat Med. 2022 March 15; 41(6): 964–980. doi:10.1002/sim.9298.

B-Value and Empirical Equivalence Bound: A New Procedure of Hypothesis Testing

Yi Zhao[*,1], **Brian S. Caffo**[2], **Joshua B. Ewen**[3]

[1]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indiana, United States

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Maryland, United States

[3]Kennedy Krieger Institute and Johns Hopkins University School of Medicine, Maryland, United States

Summary

In this study, we propose a two-stage procedure for hypothesis testing, where the first stage is conventional hypothesis testing and the second is an equivalence testing procedure using an introduced Empirical Equivalence Bound. In 2016, the American Statistical Association released a policy statement on *P*-values to clarify the proper use and interpretation in response to the criticism of *reproducibility* and *replicability* in scientific findings. A recent solution to improve reproducibility and transparency in statistical hypothesis testing is to integrate *P*-values (or confidence intervals) with practical or scientific significance. Similar ideas have been proposed via the equivalence test, where the goal is to infer equality under a presumption (null) of inequality of parameters. However, the definition of scientific significance/equivalence can sometimes be ill-justified and subjective. To circumvent this drawback, we introduce the *B*-value and the Empirical Equivalence Bound, which are both estimated from the data. Performing a second-stage equivalence test, our procedure offers an opportunity to improve the reproducibility of findings across studies.

Keywords

Empirical equivalence bound; Equivalence test; Hypothesis testing

1 | INTRODUCTION

Confidence intervals (CIs) are widely used for statistical inference. In a traditional two-sample comparative study, one uses a $100(1-\alpha)\%$ confidence interval of the difference to draw an inference. Via a hypothesis test of equal means (say), when the $100(1-\alpha)\%$ CI does not cover zero, it is concluded that there exists evidence of a difference between the

*Corresponding author. yz125@iu.edu.
Present Address
410 10th Street, Indianapolis, IN 46202.

two groups, controlling the Type I error rate at α. When the CI covers zero, it is concluded that there is insufficient evidence to suggest a difference. However, as is often emphasized in introductory courses, this insufficient evidence of difference does not imply equivalence, absence of evidence is not evidence of absence, so to speak. To test for equivalence, an equivalence test is performed where the typical roles of the null and alternative are reversed and explicit bounds implying equivalence are specified. It should be emphasized that the equivalence margin should be determined before data collection[1]. If the equivalence test is implemented in a *post hoc* sense, i.e. if the confidence interval was verified to include zero prior to investigating equivalence, special care must be taken to avoid erroneous conclusions[2].

Equivalence test procedures have long been studied to examine the equivalence of two drug formulations. Westlake[3,4] proposed the use of symmetric confidence intervals in lieu of conventional confidence intervals in bioequivalence trials. These symmetric confidence intervals decrease the effective length of the interval while increasing the confidence coefficient. Here, the effective length of a confidence interval $[L, U]$ is not $U - L$, but rather $2 \max\{|L|, |U|\}$. Anderson and Hauck[5] and Hauck and Anderson[6] introduced *t*-test procedures that were shown to be more powerful than the symmetric/shortest confidence interval approach when testing for equivalence. Schuirmann[7] considered a two one-sided tests procedure. Compared with the *power* method by Hauck and Anderson[6], the two one-sided tests procedure demonstrates superior properties. Hybridizing the power method and the two one-sided tests procedure, Munk[8] corrected the inflated Type I error rate in the power method. Liu[9], Hsu et al.[10], and Seaman and Serlin[11] proposed the use of $100(1 - 2\alpha)\%$ CIs to construct the so-called *equivalence confidence interval*, which is also symmetric about zero, and showed that the equivalence confidence interval can lead to a more powerful test, as the effective length of the interval is smaller. Seaman and Serlin[11] also suggested a sequential method to assess the difference between two means. The method starts with a conventional two-sample *t*-test. If the null hypothesis cannot be rejected, one can proceed to the equivalence test comparing the equivalence confidence interval with the nominal equivalence interval. For other extensions to the equivalence test procedure, one can refer to the discussion in Seaman and Serlin[11]. When the significance of the *t*-test is not obtained, various studies advocate for a post-experiment power calculation. However, this approach sits on an inappropriate statistical hypothesis[6,7] and suffers from fatal logical flaws, as discussed in Hoening and Heisey[2]. As an alternative, a confidence interval or equivalence test was suggested in Hoening and Heisey[2]. Recently, the equivalence testing has been generalized not limited to testing two means, such as assessing structural equation models[12]. Similarly based on an equivalence range around a null hypothesis of values, Blume et al.[13] introduced an extension of the *p*-value, called the second generation *p*-value, to represent "the proportion of data-supported hypotheses that are also null hypotheses". Lakens and Delacre[14] contrasted the second generation *p*-value with equivalence testing concluding that results were similar under optimal conditions, but prefer equivalence testing under suboptimal cases for interpretability.

One critical concern about equivalence tests (including the second generation *p*-value) is the choice of equivalence bounds. In an equivalence study, one rejects the null hypothesis of inequivalence if a $100(1 - 2\alpha)\%$ interval is entirely contained within the pre-specified

equivalence bounds. The bounds, typically symmetric around zero, are chosen to represent a trivial difference so that a true difference less than the bounds is considered equivalent. For example, Figure 1 presents possible conclusions with different choices of bounds. Of course, with narrow bounds, the equivalence test is less likely to conclude equivalence. In recent practical guidance, it is suggested to set the equivalence bounds based on standardized effect sizes taking practical feasibility into consideration or otherwise setting the bounds based on benchmarks for small, medium, and large effects[15]. Still, the choices can be subjective and lack principled considerations.

In this study, we introduce the concepts of the B-value and the Empirical Equivalence Bound: the minimum equivalence bound in principle that leads to equivalence when equivalence is true. The B-value is analogous to the attained significance level interpretation of a P-value. That is, the attained significance level is the smallest type I error rate for which one would reject the null hypothesis. The B-value is the smallest symmetric equivalence bound for which one would reject in a test of equivalence. This is inherently useful when one wants to test equivalence, but does not have a natural bound to work with. By reporting the B-value, the reader can easily apply whatever bound they choose, not unlike how a reader can easily employ any error rate they choose if P-values are reported. Here, we would like to note that when an equivalence bound is well-justified as clinically meaningful before data collection, one should use this pre-defined equivalence interval to test for equivalence. Our proposal should only be used when no such equivalence bound is available or for reporting purposes where a reader may be interested in applying a different bound. In addition, as a univariate summary, the statistical properties of the B-value are easier to derive, which we exploit below.

We follow the sequential method recommended by Seaman and Serlin[11], while further studying the properties of the equivalence confidence interval and addressing the issue of how to determine a nominal equivalence bound. In Section 2, we introduce the concept of B-value, which is defined as the maximum magnitude of the $100(1-2\alpha)\%$ CI bounds. The B-value is the smallest symmetric equivalence bound for which one would conclude equivalence. We derive the distribution of the B-value, as well as the conditional distribution based on the hypothesis testing result in the conventional two-sample t-test. Based on these distributions, we then introduce the Empirical Equivalence Bound (EEB), which can be used for equivalence tests. A two-stage testing procedure to compare two group means is suggested. A toy example and a simulation study are included to demonstrate the implementation and to evaluate the performance. This data-driven procedure requires no prior knowledge as to what level the two groups are equivalent. In Section 3, we apply our proposed procedure to the Iris Data[16] available in the open-source software R[17] and an in-house dataset from the study of children with autism spectrum disorder (ASD). Section 4 gives a summary and discussion.

2 | B-VALUE AND THE EMPIRICAL EQUIVALENCE BOUND

2.1 | Formulation

Consider a two-sample t-test setting with hypotheses

$$\mathrm{H}_0: \delta = 0 \quad \text{versus} \quad \mathrm{H}_1: \delta \neq 0, \tag{1}$$

where $\delta = \mu_1 - \mu_2$ is the difference of two population averages. A standard procedure for performing hypothesis testing is to construct the confidence interval. Let $[L_0, U_0]$ denote the $100(1-\alpha)\%$ confidence interval, where

$$L_0 = \hat{\delta} - t_{v,1-\alpha/2}S, \quad U_0 = \hat{\delta} + t_{v,1-\alpha/2}S,$$

$\hat{\delta} = \bar{x}_1 - \bar{x}_2$ is an estimate of δ with $\bar{x}_1$ and $\bar{x}_2$ as the sample average of the two groups; $t_{v,1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of a t-distribution with degrees of freedom v; S is the pooled standard error under the assumption of constant variances across groups:

$$S = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \times \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}, \tag{2}$$

where S_1^2 and S_2^2 are the sample variance of the two groups, and n_1 and n_2 are the sample sizes. Hypothesis test (1) is based on whether $[L_0, U_0]$ covers zero. However, when $0 \in [L_0, U_0]$, via normal testing logic, one cannot directly conclude equivalence of the two groups.

As suggested in Seaman and Serlin[11], in order to evaluate equivalence, an equivalence test can be conducted. Later, we discuss the implications of testing hypothesis (1) followed by performing an equivalence test. Two types of equivalence test are discussed, one is based on the marginal result, regardless of the conclusion of testing (1), and the other is based on the conditional result which is conditional on whether hypothesis (1) is rejected or not. For now, assume the equivalence test is the only test performed.

In equivalence testing, one is testing the hypotheses

$$\mathrm{H}_3: |\delta| \geq \Delta \quad \text{versus} \quad \mathrm{H}_4: |\delta| < \Delta \tag{3}$$

where Δ is a pre-specified equivalence bound. The alternative hypothesis, $|\delta| < \Delta$ represents equivalence in the sense that Δ is chosen to represent a trivially small difference given the context. Here, instead of using the $100(1-\alpha)\%$ confidence interval, a $100(1-2\alpha)\%$ confidence interval is formulated[11], denoted as $[L, U]$, where

$$L = \hat{\delta} - t_{v,1-\alpha}S, \quad U = \hat{\delta} + t_{v,1-\alpha}S, \tag{4}$$

and $t_{v,1-\alpha}$ is the $100(1-\alpha)\%$ quantile of a t-distribution with degrees of freedom v. The classic equivalence test is to compare this interval with predetermined equivalence bounds. If the interval lies entirely within the bounds, the null hypothesis is rejected (equivalence concluded, see Figure 1). Otherwise, the null hypothesis is not rejected and there is insufficient evidence to conclude equivalence. It is important to emphasize that the conclusion is subject to the choice of Δ. In this study, we propose a procedure in which the

equivalence bound is derived from the data, and we call it the *Empirical Equivalence Bound* (EEB).

To obtain the EEB, we first introduce the *B-value*, $B = \max\{|L|, |U|\}$. If one takes nothing else from this manuscript, consider that the B-value is useful to report in the sense that if $B < \Delta$ one rejects and concludes equivalence. This is analogous to the attained significance level where one rejects if $P < \alpha$. Since it is common to want to test equivalence, but not be in possession of meangingful bounds, reporting B is useful as a reader can then perform the test on whatever bounds they believe are most relevant.

In this manuscript we investigate the properties of B. Specifically, we derive the distribution of B, as well as the conditional distribution of B given the test result of the first-step hypothesis testing (1); that is, conditional on $0 \in [L_0, U_0]$ or $0 \notin [L_0, U_0]$. This is useful as it is common to investigate equivalence after failing to reject classic hypotheses. However, it is well understood that ignoring the first rejection leads to erroneous conclusions[11,2].

We also consider the distribution of B given $0 \notin [L_0, U_0]$. That is, considering testing equivalence, hypotheses (3), after having rejected the null hypothesis from (1). This is done primarily for completeness. However, it's possible that a researcher might want to investigate potential triviality of a rejection of the traditional hypothesis test.

The most related study is Seaman and Serlin[11], where $[-B, B]$ was referred to as the *equivalence confidence interval* which was recommended for use when the effect is small. In the following Result 1, we focus on the scenario when the true δ is zero, exact equivalence. In Section A of the supplementary material, we provide the (conditional) distribution of B under the general scenario of $\delta \in \mathbb{R}$.

Result 1.—For a two-sample t-test as in (1), when the true $\delta = 0$,

1. the cumulative distribution function of B is:

$$F_B(b \mid \mathrm{H}_0) = \begin{cases} 0 & \text{if } b < St_{v,\,1-\alpha} \\ 2F_t\left(b/S - t_{v,\,1-\alpha}; v\right) - 1 & \text{if } b \geq St_{v,\,1-\alpha} \end{cases};$$

2. the conditional cumulative distribution function of B given $0 \in [L_0, U_0]$ is:

$$F_B(b \mid 0 \in [L_0, U_0], \mathrm{H}_0) =$$

$$\begin{cases} 0 & \text{if } b < St_{v,\,1-\alpha} \\ \left\{2F_t\left(b/S - t_{v,\,1-\alpha}; v\right) - 1\right\}/(1-\alpha) & \text{if } St_{v,\,1-\alpha} \leq b < S\left(t_{v,\,1-\alpha} + t_{v,\,1-\alpha/2}\right); \\ 1 & \text{if } b \geq S\left(t_{v,\,1-\alpha} + t_{v,\,1-\alpha/2}\right) \end{cases}$$

3. the conditional cumulative distribution function of B given $0 \notin [L_0, U_0]$ is:

$$F_B(b \mid 0 \notin [L_0, U_0], \mathrm{H}_0) = \begin{cases} 0 & \text{if } b < S\left(t_{v,1-\alpha} + t_{v,1-\alpha/2}\right) \\ \left\{F_t\left(b/S - t_{v,1-\alpha}; v\right) - (1-\alpha/2)\right\}/(\alpha/2) & \text{if } b \geq S\left(t_{v,1-\alpha} + t_{v,1-\alpha/2}\right) \end{cases},$$

where $F_t(\cdot;\ v)$ is the cumulative distribution function of a Student's t-distribution with degrees of freedom v.

The *Empirical Equivalence Bound* is defined as the bound of an equivalence test such that when the true population difference is exactly zero, an equivalence test rejects with probability β.

Definition 1 (Empirical Equivalence Bound).—Assume that the parameter of interest is the difference in the population means, $\delta = \mu_1 - \mu_2$. Consider a hypothesis testing problem $\mathrm{H}_0 : \delta = 0$ of level α. For a given $\beta \in (0, 1)$, under test result C, the Empirical Equivalence Bound at level β is defined as:

$$\mathrm{EEB}_\alpha(\beta \mid C) = \inf_{b \in [0,\infty]} \{b :\ F_B(b \mid C, \mathrm{H}_0) \geq \beta\}. \tag{5}$$

Here, $C \in \{\varnothing, 0 \in [L_0, U_0], 0 \notin [L_0, U_0]\}$ denotes the status of the hypothesis test, and $F_B(\cdot \mid C, \mathrm{H}_0)$ is the conditional cumulative distribution function of B. If the test result is unknown, $C = \varnothing$, and $F_B(\cdot \mid \varnothing, \mathrm{H}_0) = F_B(\cdot \mid \mathrm{H}_0)$ is the marginal distribution of B.

Under the condition $0 \in [L_0, U_0]$ or $0 \notin [L_0, U_0]$, the $\mathrm{EEB}(\beta)$ defined in Definition 1 has the following explicit form:

$$\mathrm{EEB}_\alpha(\beta \mid 0 \in [L_0, U_0]) = S\left\{F_t^{-1}\left(\frac{\beta(1-\alpha)+1}{2}; v\right) + t_{v,1-\alpha}\right\},$$

$$\mathrm{EEB}_\alpha(\beta \mid 0 \notin [L_0, U_0]) = S\left\{F_t^{-1}\left(1 - \frac{\alpha(1-\beta)}{2}; v\right) + t_{v,1-\alpha}\right\},$$

where $F_t^{-1}(\cdot\,; v)$ is the inverse cumulative distribution function of a Student's t-distribution with degrees of freedom v.

2.2 | Properties of the EEB

The EEB has the following properties:

- **i.** For fixed α, $\mathrm{EEB}_\alpha(\beta|C)$ is a non-decreasing function of β;
- **ii.** For fixed β, $\mathrm{EEB}_\alpha(\beta|C)$ is a non-increasing function of α.

The proof of these two properties is straightforward following the fact that the cumulative density function is right continuous and non-decreasing. With the same significance level in

the first-step t-test, it requires the equivalence interval to be wider for a higher confidence in the second-step equivalence test. On the other hand, when the level in the second-step equivalence test is fixed, a higher confidence level in the first step (lower α) demands a wider equivalence interval as well.

The following proposition presents the relationship between the three cumulative density functions in Result 1 and the correspondingly defined EEB.

Proposition 1.—Consider the three distributions of B Result 1. For any b, we have

$$F_B(b \mid 0 \notin [L_0, U_0], \mathrm{H}_0) \leq F_B(b \mid \mathrm{H}_0) \leq F_B(b \mid 0 \in [L_0, U_0], \mathrm{H}_0).$$

Therefore, for fixed α and β, the EEBs have the following relationship

$$\mathrm{EEB}_\alpha(\beta \mid 0 \notin [L_0, U_0]) \geq \mathrm{EEB}_\alpha(\beta) \geq \mathrm{EEB}_\alpha(\beta \mid 0 \in [L_0, U_0]),$$

where $\mathrm{EEB}_\alpha(\beta) = \mathrm{EEB}_\alpha(\beta \mid \varnothing)$.

Proposition 1 demonstrates that, given the result from the two-sample t-test, the conditional EEB appropriately shrinks/expands the equivalence interval in the equivalence test. For example, if the two-sample t-test does not reject the null, the B-value, as well as the conditional EEB, shrinks toward zero, compared to the values not conditioning on the results of Step 1 testing. If the test rejects the null even though the true parameter is zero, then a wider equivalence interval is required to correct for this false positive, which is achieved with a greater EEB value. In other words, the EEB allows one to interpret equivalence post testing.

2.3 | A two-stage testing procedure

Using the defined EEB, we propose a two-stage testing procedure when comparing two means. The procedure is summarized in Figure 2. The first stage is the conventional two-sample t-test. Based on whether the $100(1-\alpha)\%$ confidence interval covers zero, we calculate the conditional EEB at level β denoted as $\Delta_\alpha^{(r)}(\beta)$, where $r = 0$ if $0 \in 100(1-\alpha)\%$ CI and $r = 1$ otherwise. The second stage compares the $100(1-2\alpha)\%$ CI (denoted by $[L, U]$) and the β-level empirical equivalence interval (denoted by $\left[-\Delta_\alpha^{(r)}(\beta), \Delta_\alpha^{(r)}(\beta)\right]$ for $r = 0, 1$). When the first stage result is $0 \in 100(1-\alpha)\%$ CI, if the $100(1-2\alpha)\%$ CI is fully contained in the empirical equivalence interval, there is sufficient evidence to conclude equivalence of the two groups. Otherwise, no confirmatory conclusion can be achieved. When the first stage result is $0 \notin 100(1-\alpha)\%$ CI, if the empirical equivalence interval covers the $100(1-2\alpha)\%$ CI, one can conclude that the two means are actually equivalent and the second-stage equivalence test corrects the false positive discovery in the first stage. If there is overlap between the two intervals, there is no confirmatory conclusion. For sufficiently high β level, if there is no overlap between the two intervals, this can be seen as a confirmation of the significant finding in the first stage.

2.4 | An example

In this section, we use simulated examples to elaborate on the definition and the properties of the B-value and the EEB and demonstrate the implementation of the two-stage testing procedure in Figure 2. Assuming that the true δ parameter is zero, and we generate two-sample data from equivalent normal distributions. The estimate of δ, standard error (S), and the $100(1-\alpha)\%$ confidence interval are then attained. By examining if the $100(1-\alpha)\%$ confidence interval covers zero, with a designated level β, we then calculate the EEB under the corresponding condition and compare it with the $100(1-2\alpha)\%$ confidence interval.

For example, with the sample sizes $n_1 = n_2 = 10$ (and thus $v = 18$), we generate data for both groups from a standard normal distribution (i.e. $\delta = 0$). We simulate data for two scenarios: (1) $0 \in [L_0, U_0]$ (Examples (i)–(ii) in Table 1) and (2) $0 \notin [L_0, U_0]$ (Examples (iii)–(v) in Table 1). Here, to compare between the two scenarios, we keep the standard error the same and generate $\hat{\delta}$ from the sample mean distribution. The statistics are presented in Table 1. Figure 3a presents the marginal distribution of the B-value, and Figures 3b and 3c present the conditional distribution, together with the EEB at various β-values. Figures 4a and 4b compare the three distributions and the EEB values which verify Proposition 1. Under Scenario (1), for a given β, the knowledge of not rejecting the null in the two-sample t-test decreases the EEB value, making it more stringent in the second-step equivalence test. Under Scenario (2), with a bigger conditional EEB, the equivalence interval is wider, so that performing an equivalence test may help rectify the false positive finding in the two-sample t-test. Therefore, conditional on the result from the two-sample t-test, the conditional EEB improves the performance of the second-step equivalence test. Figure 4 also demonstrates one property of the EEB that for a fixed significance level α, the EEB is a non-decreasing function of β.

Setting $\beta = 0.95$, Figure 5 demonstrates the two-step testing procedure and the conclusions. For Examples (i)–(ii), the first-step conventional two-sample t-test concludes that no sufficient evidence supports the difference between the two groups. The test is followed by a second-step test using the EEB with $\beta = 0.95$. Example (i) yields the conclusion that conditional on the result from the first step, setting the equivalence margin at 1.131, the probability of equivalence is at least 95%. However, for Example (ii), the 90% confidence interval is not completely covered by the equivalence interval. Thus, one cannot conclude equivalence. For Examples (iii)–(v), the Step 1 test concludes a significant difference between the two groups. However, after Step 2, it concludes equivalence in Example (iii) at the level of $\beta = 0.95$. With the same equivalence margin, we can only conclude the difference for Example (v).

2.5 | Comparison between the marginal EEB and conditional EEB

In this section, we use a simulation study to demonstrate the difference in the performance between the marginal equivalence bound and the conditional equivalence bound. We use the same standard error and sample sizes in Section 2.4 for data generation, We first generate a random variable from the t-distribution with degrees of freedom $v = 18$, which is the t-statistic in the two-sample t-test. Using the generated t-statistic and the standard error, we derive the estimated difference, $\hat{\delta}$, and the 95% and 90% confidence intervals, For the Step

2 test, we set $\beta = 0.95$. Using the standard error, the marginal and conditional empirical equivalence bound (EEB) are calculated. For the marginal EEB, it is directly used for the inference of equivalence; and for the conditional EEB, based on the result of $0 \notin [L_0, U_0]$ or $0 \in [L_0, U_0]$, the appropriate conditional EEB is used for the inference of equivalence, The simulation is repeated for 10000 repetitions, Among these 10000 repetitions, 4.8% are rejected in Step 1 test. Table 2 presents the proportion of concluding equivalence in Step 2. For both types of EEB, the proportion of concluding equivalence is approximately 0.952, which is expected as we set $\beta = 0.95$. We also calculate the proportion of concluding equivalence for each type of conclusion in Step 1. Using the conditional EEB, the proportion of equivalence for each type of conclusion is around 0.95. However, using the marginal EEB, the proportion of equivalence of not rejecting the null in Step 1 is 100% and the proportion of equivalence of rejecting the null in Step 1 is 0%. This indicates that using the marginal EEB, when the null hypothesis is falsely rejected, having Step 2 testing will not help rectify the conclusion. Thus, using the conditional EEB is the appropriate choice.

2.6 | Generalization to *z*-test

Many parameter estimators follow a Gaussian distribution asymptotically, such as method of moments and maximum likelihood estimators. For these estimators, hypothesis testing can be conducted through a *z*-test, where the *z*-score is calculated based on the asymptotic distribution. For a *z*-test, one can replace the *t*-distribution quantiles and cumulative distribution function in Section 2.1 with the corresponding quantities from the standard normal distribution and all the results follow.

3 | DATA APPLICATIONS

We demonstrate the implementation of the proposed two-stage testing approach via two datasets, one publicly available and one in-house. The utilization of the empirical equivalence bound (EEB) is under the scenario that the equivalence bound is not well-defined and testing for equivalence is part of the study interest. When an equivalence margin is well-defined, direct utilization of such margin in the equivalence testing is preferred. In both applications, we present the conclusions with $\beta = 0.95$, analogous to the 95% confidence level.

3.1 | Iris Data

We first demonstrate the practical implementation using the Iris Data Set available in the open-source software R[17]. The data were collected by Edgar Anderson to compare between three related species of irises, namely setosa, versicolor, and virginica[16]. There are 50 samples of each species. In this section, we present the comparison of two measures, sepal length and sepal width. Figures 6a and 6c present the scatter plot of each species and the 95% confidence interval, where the confidence intervals are derived from an analysis of variance (ANOVA) model. Table 3 presents the statistics of the pair-wise comparisons. From the table, for both the comparison of sepal length and sepal width, the first-step test suggests a significant mutual difference between the three species, where all the significance remains after performing Tukey's test. Based on the significant results, we further conduct an equivalence test using the conditional EEB with $\beta = 0.95$. Figures 6b and 6d present

the results. After the second stage of testing, the difference between setosa and versicolor and the difference between setosa and virginica are significant in both sepal length and sepal width. For the comparison between versicolor and virginica, we cannot conclude either difference or equivalence in sepal length, as the empirical equivalence interval and the 90% confidence interval are partly overlapped; and we conclude equivalence in sepal width as the 90% confidence interval is completely covered by the equivalence interval. The equivalence of versicolor and virginica in sepal width is more consistent with the clustering result, where one cluster contains setosa and the other cluster contains versicolor and virginica[18].

3.2 | Learning of Skilled Movements via Imitation in ASD

In this section, we present the implementation to an in-house dataset collected in a study of children with autism spectrum disorder (ASD). Children with ASD demonstrate behavior deficits in various domains, including social, language, and motor skills. In this study, we focus on the performance of gesture learning via imitation as it requires both motor and cognitive skills learned from interactions with the environment and abnormality is highly prevalent in ASD. In addition, it is easier to study in a laboratory context than other, such as social and communicative, skills. Eighteen participants with ASD (IQ > 80, ages 8–12.9 years) and 19 typically developing (TD) peers were recruited. A novel-gesture-learning task previously studied in neurotypical adults was employed. In each trial, the participant watched a video of the right arm performing a novel, meaningless gesture and then performed the gesture. Feedback was given by a research assistant. The performance was measured by the total number of gestures scored, where correctness was judged if all elements were performed and in the correct order. There were a total of 12 gestures repeated for at least four repetitions. More details about the study can be found in McAuliffe et al.[19]. With repeated measures, we analyzed the data using a mixed-effects model with group, repetition, and their interaction as the predictors to compare the performance between groups and examine the variations across repetitions. We also included age, sex, and motor coordination quantified via the Physical and Neurological Assessment of Subtle Signs (PANESS) to remove the potential confounding effect. Figure 7 shows the scatter plot of the data and the 95% confidence intervals derived from the mixed-effects model. Table 4 presents the statistics of group comparison for each repetition. From the table, a significant difference between the two groups appears at repetitions 2 and 3 and the performance reconverges by repetition 4. In the second step of equivalence testing, with $\beta = 0.95$, one can conclude equivalence at repetitions 1, 2, and 3, while not for repetition 4. The study hypothesis was that "children with ASD show alterations in learning novel gestures via imitation". Following the proposed two-stage procedure, the current data do not offer sufficient evidence to support the hypothesis.

4 | DISCUSSION

The use of *P*-values in hypothesis testing has a long history, dating back to 1925 when R. A. Fisher introduced and promoted it for rejecting a null hypothesis when small[20]: "We shall not often be astray if we draw a conventional line at 0.05." Since then, there are ongoing discussions about how to correctly use and interpret them. Though alternative statistics, for example, confidence intervals, effect sizes, or Bayes factors, are available, it has been shown

that the interpretation of uncertainty is similar[21]. Common misunderstandings and misuse of *P*-values are likely partially responsible for general confusion and mistrust of empirical findings. In 2015, the editors of Basic and Applied Social Psychology (BASP) decided to ban *P*-values null hypothesis significance testing,[22], a controversial move, even among opponents of null hypothesis significance testing and *P*-value usage. In 2016, the American Statistical Association announced a policy statement on *P*-values[23]. In the statement, the authors noted that "the statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions". In an echo to Peng[24] and Leek and Peng[25], the statement also accentuated that "misunderstanding and misuse of statistical inference is only one cause of the reproducibility crisis". Hashing out opinions from more than two dozen well-respected statisticians, the statement outlines six principles in order to regulate the proper use and interpretation of the *P*-values. One principle states that "a *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result", as a *P*-value highly depends on the precision of the estimate (or the sample size). As discussed, recently, Blume et al.[13] introduced the second-generation *P*-value and Goodman et al. [26] proposed a hybrid effect size plus *P*-value criterion. Both criteria assemble the *P*-value (or confidence interval) with a practical/scientific significance in the testing procedure. However, same as in the equivalence test, the definition of practical/scientific significance can sometimes be subjective and arbitrary.

In this study, we introduced the B-value and the Empirical Equivalence Bound and proposed a two-stage procedure when comparing two means from Gaussian distributed data. Our method is a data-driven procedure relaxing the knowledge of the equivalence level in an equivalence test. In the equivalence test, the conclusion highly depends on the nominal equivalence bound, which can sometimes be ill-justified and subjective. Our method eliminates this drawback by using the empirical equivalence bound derived from the data. On the other hand, performing a second-stage equivalence test also provides an opportunity to examine whether the significant result in the conventional two-sample *t*-test is a false positive discovery. This new two-stage testing procedure may then help improve the reproducibility of findings across studies.

ACKNOWLEDGMENTS

Yi Zhao was partially supported by NIH grants P01HL158507, P30AG072976, and U54AG065181. Brian Caffo was partially supported by NIH grants R01EB029977, P50HD103538, and U54DA049110. Joshua Ewen was partially supported by NIH grant R21NS091569.

DATA AVAILABILITY STATEMENT

The Iris Dataset in our application is available in the open-source software R[17]. The ASD data are available to qualified researchers through NIH's National Database for Autism Research. The R package Bvalue is available on CRAN (https://CRAN.R-project.org/package=Bvalue) for implementation.

APPENDIX

A A DISTRIBUTION OF THE *B*-VALUE

In this section, we derive the density function of B. Result 1 is a special scenario when the true parameter of interest δ is zero.

First, we derive the marginal cumulative distribution function of B. $B = \max\{|L|, |U|\}$, where

$$L = \hat{\delta} - t_{v,1-\alpha}S, \quad U = \hat{\delta} + t_{v,1-\alpha}S,$$

and the estimator $\hat{\delta}$ follows a Student's t-distribution with degrees of freedom v

$$\frac{\hat{\delta} - \delta}{S} \sim t(v),$$

where S is the standard error and $t_{v,1-a}$ is the $100(1-a)\%$ quantile of a t-distribution with degrees of freedom v. Thus, $[L, U]$ is the $100(1-2a)\%$ confidence interval.

For $\forall\ b \geq 0$, the cumulative distribution function of B is

$$\begin{aligned}
F_B(b) &= \mathbb{P}(B \leq b) \\
&= \mathbb{P}(|L| \leq b, B = |L|) + \mathbb{P}(|U| \leq b, B = |U|) \\
&= \mathbb{P}(|L| \leq b, L < -U < 0 < U) + \mathbb{P}(|L| \leq b, L < U < 0) \\
&\quad + \mathbb{P}(|U| \leq b, L < 0 < -L < U) + \mathbb{P}(|U| \leq b, 0 < L < U) \\
&= \mathbb{P}(-L \leq b, L < -U < 0 < U) + \mathbb{P}(-L \leq b, L < U < 0) \\
&\quad + \mathbb{P}(U \leq b, L < 0 < -L < U) + \mathbb{P}(U \leq b, 0 < L < U).
\end{aligned}$$

Denote $F_t(\cdot;\ v)$ as the cumulative distribution function of a Student's t-distribution with degrees of freedom v,

$$\begin{aligned}
\mathbb{P}(|L| \leq b, L < -U < 0 < U) &= \mathbb{P}\left(\frac{\hat{\delta} - \delta}{S} \geq t_{v,1-\alpha} + \frac{-b-\delta}{S}, \frac{\hat{\delta} - \delta}{S} < -\frac{\delta}{S}, \frac{\hat{\delta} - \delta}{S} > -t_{v,1-\alpha} - \frac{\delta}{S}\right) \\
&= \mathbb{P}\left(\max\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}, -t_{v,1-\alpha} - \frac{\delta}{S}\right) \leq \frac{\hat{\delta} - \delta}{S} < -\frac{\delta}{S}\right);
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(|L| \leq b, L < U < 0) &= \mathbb{P}\left(\frac{\hat{\delta} - \delta}{S} \geq t_{v,1-\alpha} + \frac{-b-\delta}{S}, \frac{\hat{\delta} - \delta}{S} < -t_{v,1-\alpha} - \frac{\delta}{S}\right) \\
&= \mathbb{P}\left(t_{v,1-\alpha} + \frac{-b-\delta}{S} \leq \frac{\hat{\delta} - \delta}{S} < -t_{v,1-\alpha} - \frac{\delta}{S}\right);
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(|U| \leq b, L < 0 < -L < U) &= \mathbb{P}\left(\frac{\hat{\delta} - \delta}{S} \leq -t_{v,1-\alpha} + \frac{b-\delta}{S}, \frac{\hat{\delta} - \delta}{S} > -\frac{\delta}{S}, \frac{\hat{\delta} - \delta}{S} < t_{v,1-\alpha} - \frac{\delta}{S}\right) \\
&= \mathbb{P}\left(-\frac{\delta}{S} < \frac{\hat{\delta} - \delta}{S} \leq \min\left(-t_{v,1-\alpha} + \frac{b-\delta}{S}, t_{v,1-\alpha} - \frac{\delta}{S}\right)\right);
\end{aligned}$$

$$\mathbb{P}(|U| \le b, 0 < L < U) = \mathbb{P}\left(\frac{\hat{\delta}-\delta}{S} \le -t_{v,1-\alpha} + \frac{b-\delta}{S}, \frac{\hat{\delta}-\delta}{S} > t_{v,1-\alpha} - \frac{\delta}{S}\right)$$
$$= \mathbb{P}\left(t_{v,1-\alpha} - \frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S} \le -t_{v,1-\alpha} + \frac{b-\delta}{S}\right).$$

Thus, if $St_{v,1-\alpha} \le b < 2St_{v,1-\alpha}$

$$F_B(b) = F_t\left(-t_{v,1-\alpha} + \frac{b-\delta}{S}; v\right) - F_t\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}; v\right);$$

and if if $b \ge 2St_{v,1-\alpha}$

$$F_B(b) = F_t\left(-t_{v,1-\alpha} + \frac{b-\delta}{S}; v\right) - F_t\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}; v\right).$$

Therefore, the cumulative function of B is for $b \ge St_{v,1-\alpha}$

$$F_B(b) = F_t\left(-t_{v,1-\alpha} + \frac{b-\delta}{S}; v\right) - F_t\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}; v\right); \tag{A1}$$

and the probability density function is

$$f_B(b) = \frac{\mathrm{d}F_B(b)}{\mathrm{d}b} = \frac{1}{S}\left\{f_t\left(-t_{v,1-\alpha} + \frac{b-\delta}{S}; v\right) + f_t\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}; v\right)\right\}, \tag{A2}$$

where $f_t(\cdot\,; v)$ is the probability density function of a Student's t-distribution with degrees of freedom v.

With a special case that $\delta = 0$, function (A1) degenerates to the distribution function in Result 1.

Now, we consider the conditional distribution of B given that $0 \in [L_0, U_0]$.

$$F_B(b \mid 0 \in [L_0, U_0]) = \mathbb{P}(B \le b \mid 0 \in [L_0, U_0]) = \frac{\mathbb{P}(B \le b, 0 \in [L_0, U_0])}{\mathbb{P}(0 \in [L_0, U_0])}.$$

$$\mathbb{P}(0 \in [L_0, U_0]) = \mathbb{P}\left(\hat{\delta} - t_{v,1-\alpha/2}S < 0 < \hat{\delta} + t_{v,1-\alpha/2}S\right)$$
$$= \mathbb{P}\left(\frac{\hat{\delta}-\delta}{S} < t_{v,1-\alpha/2} - \frac{\delta}{S}, -t_{v,1-\alpha 2} - \frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S}\right)$$
$$= F_t\left(t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right) - F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right).$$

$$
\begin{aligned}
&\mathbb{P}(B \le b, 0 \in [L_0, U_0]) \\
&= \mathbb{P}[B \le b, -U_0 < L_0 < 0 < U_0] + \mathbb{P}[B \le b, L_0 < 0 < U_0 < -L_0] \\
&= \mathbb{P}\left[U \le b, -\hat{\delta} - t_{v,1-\alpha/2}S < \hat{\delta} - t_{v,1-\alpha/2}S < 0 < \hat{\delta} + t_{v,1-\alpha/2}S\right] \\
&\quad + \mathbb{P}\left[L \le b, \hat{\delta} - t_{v,1-\alpha/2}S < 0 < \hat{\delta} + t_{v,1-\alpha/2}S < -\hat{\delta} + t_{v,1-\alpha/2}S\right] \\
&= \mathbb{P}\left[\hat{\delta} + t_{v,1-\alpha}S \le b, \hat{\delta} > 0, \hat{\delta} - t_{v,1-\alpha/2}S < 0 < \hat{\delta} + t_{v,1-\alpha/2}S\right] \\
&\quad + \mathbb{P}\left[-\hat{\delta} + t_{v,1-\alpha}S \le b, \hat{\delta} < 0, \hat{\delta} - t_{v,1-\alpha/2}S < 0 < \hat{\delta} + t_{v,1-\alpha/2}S\right] \\
&= \mathbb{P}\left[\frac{\hat{\delta}-\delta}{S} \le \frac{b-\delta}{S} - t_{v,1-\alpha}, \frac{\hat{\delta}-\delta}{S} > -\frac{\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S} < t_{v,1-\alpha/2} - \frac{\delta}{S}\right] \\
&\quad + \mathbb{P}\left[\frac{\hat{\delta}-\delta}{S} \ge t_{v,1-\alpha} + \frac{-b-\delta}{S}, \frac{\hat{\delta}-\delta}{S} < -\frac{\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S} < t_{v,1-\alpha/2} - \frac{\delta}{S}\right] \\
&= \mathbb{P}\left[-\frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S} \le \min\left(\frac{b-\delta}{S} - t_{v,1-\alpha}, t_{v,1-\alpha/2} - \frac{\delta}{S}\right)\right] \\
&\quad + \mathbb{P}\left[\max\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S}\right) < \frac{\hat{\delta}-\delta}{S} < -\frac{\delta}{S}\right] \\
&= F_t\left\{\min\left(\frac{b-\delta}{S} - t_{v,1-\alpha}, t_{v,1-\alpha/2} - \frac{\delta}{S}\right); v\right\} - F_t\left\{\max\left(t_{v,1-\alpha} - \frac{b+\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S}\right); v\right\}.
\end{aligned}
$$

$\Rightarrow$

$$
\begin{aligned}
&F_B(b \mid 0 \in [L_0, U_0]) \\
&= \frac{F_t\left\{\min\left(\frac{b-\delta}{S} - t_{v,1-\alpha}, t_{v,1-\alpha/2} - \frac{\delta}{S}\right); v\right\} - F_t\left\{\max\left(t_{v,1-\alpha} - \frac{b+\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S}\right); v\right\}}{F_t\left(t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right) - F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right)}.
\end{aligned}
$$

- If $b \le S\left(t_{v,1-\alpha} + t_{v,1-\alpha/2}\right)$,

$$
\begin{cases}
\min\left(\frac{b-\delta}{S} - t_{v,1-\alpha}, t_{v,1-\alpha/2} - \frac{\delta}{S}\right) = \frac{b-\delta}{S} - t_{v,1-\alpha} \\
\max\left(t_{v,1-\alpha} - \frac{b+\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S}\right) = t_{v,1-\alpha} - \frac{b+\delta}{S}
\end{cases},
$$

$$
\frac{b-\delta}{S} - t_{v,1-\alpha} \ge t_{v,1-\alpha} - \frac{b+\delta}{S} \quad \Rightarrow \quad b \ge St_{v,1-\alpha};
$$

- if $b > S\left(t_{v,1-\alpha} + t_{v,1-\alpha/2}\right)$,

$$
\begin{cases}
\min\left(\frac{b-\delta}{S} - t_{v,1-\alpha}, t_{v,1-\alpha/2} - \frac{\delta}{S}\right) = t_{v,1-\alpha/2} - \frac{\delta}{S} \\
\max\left(t_{v,1-\alpha} - \frac{b+\delta}{S}, -t_{v,1-\alpha/2} - \frac{\delta}{S}\right) = -t_{v,1-\alpha/2} - \frac{\delta}{S}
\end{cases},
$$

$$
t_{v,1-\alpha/2} - \frac{\delta}{S} > -t_{v,1-\alpha/2} - \frac{\delta}{S} \quad (\Rightarrow \quad F_B(b|0 \in [L_0, U_0]) = 1).
$$

Therefore,

- $b \in \left(-\infty, St_{v,1-\alpha}\right)$,

$$
F_B(b \mid 0 \in [L_0, U_0]) = 0;
$$

- $b \in \left[St_{v,1-\alpha}, S\left(t_{v,1-\alpha} + t_{v,1-\alpha/2}\right)\right]$,

$$F_B(b \mid 0 \in [L_0, U_0]) = \frac{F_t\big((b-\delta)/S - t_{v,1-\alpha}; v\big) - F_t\big(t_{v,1-\alpha} - (b+\delta)/S; v\big)}{F_t\big(t_{v,1-\alpha/2} - \delta/S; v\big) - F_t\big(-t_{v,1-\alpha/2} - \delta/S; v\big)};$$

- $b \in \big(S(t_{v,1-\alpha} + t_{v,1-\alpha 2}), \infty\big)$,

$$F_B(b \mid 0 \in [L_0, U_0]) = 1.$$

Analogously, for the conditional distribution of B given that $0 \notin [L_0,\ U_0]$,

$$F_B(b \mid 0 \notin [L_0, U_0]) = \mathbb{P}(B \le b \mid 0 \notin [L_0, U_0]) = \frac{\mathbb{P}(B \le b, 0 \notin [L_0, U_0])}{\mathbb{P}(0 \notin [L_0, U_0])}.$$

$$\begin{aligned}
\mathbb{P}(0 \notin [L_0, U_0]) &= \mathbb{P}(0 < L_0 < U_0) + \mathbb{P}(L_0 < U_0 < 0) \\
&= \mathbb{P}\big(0 < \hat{\delta} - t_{v,1-\alpha/2}S\big) + \mathbb{P}\big(\hat{\delta} + t_{v,1-\alpha/2}S < 0\big) \\
&= \mathbb{P}\left(t_{v,1-\alpha/2} - \frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S}\right) + \mathbb{P}\left(\frac{\hat{\delta}-\delta}{S} < -t_{v,1-\alpha/2} - \frac{\delta}{S}\right) \\
&= \left[1 - F_t\left(t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right)\right] + F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right) \\
&= F_t\left(\frac{\delta}{S} - t_{v,1-\alpha/2}; v\right) + F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right).
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P}(B \le b, 0 \notin [L_0, U_0]) \\
&= \mathbb{P}(B \le b, 0 < L_0 < U_0) + \mathbb{P}(B \le b, L_0 < U_0 < 0) \\
&= \mathbb{P}(U \le b, 0 < L_0 < U_0) + \mathbb{P}(-L \le b, L_0 < U_0 < 0) \\
&= \mathbb{P}\big(\hat{\delta} + t_{v,1-\alpha}S \le b, \hat{\delta} - t_{v,1-\alpha/2}S > 0\big) + \mathbb{P}\big(-\hat{\delta} + t_{v,1-\alpha}S \le b, \hat{\delta} + t_{v,1-\alpha/2}S < 0\big) \\
&= \mathbb{P}\left(t_{v,1-\alpha/2} - \frac{\delta}{S} < \frac{\hat{\delta}-\delta}{S} \le \frac{b-\delta}{S} - t_{v,1-\alpha}\right) + \mathbb{P}\left(t_{v,1-\alpha} + \frac{-b-\delta}{S} \le \frac{\hat{\delta}-\delta}{S} < -t_{v,1-\alpha/2} - \frac{\delta}{S}\right) \\
&= \left[F_t\left(\frac{b-\delta}{S} - t_{v,1-\alpha}; v\right) - F_t\left(t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right)\right] + \left[F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right) - F_t\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}; v\right)\right].
\end{aligned}$$

$\Rightarrow$

- $b \in \big(-\infty, S(t_{v,1-\alpha} + t_{v,1-\alpha/2})\big)$,

$$F_B(b \mid 0 \notin [L_0, U_0]) = 0;$$

- $b \in \big[S(t_{v,1-\alpha} + t_{v,1-\alpha 2}), \infty\big)$,

$$\begin{aligned}
&F_B(b \mid 0 \notin [L_0, U_0]) \\
&= \frac{\left[F_t\left(\frac{b-\delta}{S} - t_{v,1-\alpha}; v\right) - F_t\left(t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right)\right] + \left[F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right) - F_t\left(t_{v,1-\alpha} + \frac{-b-\delta}{S}; v\right)\right]}{F_t\left(\frac{\delta}{S} - t_{v,1-\alpha/2}; v\right) + F_t\left(-t_{v,1-\alpha/2} - \frac{\delta}{S}; v\right)}.
\end{aligned}$$

When $\delta = 0$, the (conditional) distribution functions have the same form as in Result 1.

B CALCULATION OF THE EMPIRICAL EQUIVALENCE BOUND

Following Result 1 and Definition 1, we can calculate the Empirical Equivalence Bound (EEB) from the data. When the true difference $\delta = 0$ and under conditions $0 \in [L_0, U_0]$ and $0 \notin [L_0, U_0]$, EEB has explicit forms as in Section 2.1. For other scenarios, a numerical solution of EEB can be obtained by using the bisection method.

References

1. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. Journal of general internal medicine 2011; 26(2): 192–196. [PubMed: 20857339]
2. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician 2001; 55(1): 19–24.
3. Westlake WJ. Use of confidence intervals in analysis of comparative bioavailability trials. Journal of Pharmaceutical Sciences 1972; 61(8): 1340–1341. [PubMed: 5050398]
4. Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. Biometrics 1976: 741–744. [PubMed: 1009222]
5. Anderson S, Hauck WW. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Communications in Statistics-Theory and Methods 1983; 12(23): 2663–2692.
6. Hauck WW, Anderson S. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. Journal of Pharmacokinetics and Biopharmaceutics 1984; 12(1): 83–91. [PubMed: 6747820]
7. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics 1987; 15(6): 657–680. [PubMed: 3450848]
8. Munk A An improvement on commonly used tests in bioequivalence assessment. Biometrics 1993: 1225–1230.
9. Liu H Confidence intervals in bioequivalence. In: ; 1990: 51–54.
10. Hsu JC, Hwang JG, Liu HK, Ruberg SJ. Confidence intervals associated with tests for bioequivalence. Biometrika 1994; 81(1): 103–114.
11. Seaman MA, Serlin RC. Equivalence confidence intervals for two-group comparisons of means.. Psychological Methods 1998; 3(4): 403.
12. Yuan KH, Chan W, Marcoulides GA, Bentler PM. Assessing structural equation models by equivalence testing with adjusted fit indexes. Structural Equation Modeling: A Multidisciplinary Journal 2016; 23(3): 319–330.
13. Blume JD, McGowan LD, Dupont WD, Greevy RA Jr. Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. PloS one 2018; 13(3): e0188299. [PubMed: 29565985]
14. Lakens D, Delacre M. Equivalence testing and the second generation P-value. 2018.
15. Lakens D Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science 2017; 8(4): 355–362. [PubMed: 28736600]
16. Anderson E The species problem in Iris. Annals of the Missouri Botanical Garden 1936; 23(3): 457–509.
17. R Core Team . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2018.
18. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics 1936; 7(2): 179–188.
19. McAuliffe D, Zhao Y, Pillai AS, et al. Learning of skilled movements via imitation in ASD. Autism Research 2020; 13(5): 777–784. [PubMed: 31876983]
20. Fisher RA. Statistical Methods for Research Workers. London: Oliver and Boyd . 1925.

21. Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. Perspectives on Psychological Science 2011; 6(3): 291–298. [PubMed: 26168519]
22. Trafimow D, Marks M. Editorial in Basic and Applied Social Pschology. Basic and Applied Social Pschology 2015; 37:1–2.
23. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. The American Statistician 2016; 70(2): 129–133.
24. Peng R The reproducibility crisis in science: A statistical counterattack. Significance 2015; 12(3): 30–32.
25. Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature News 2015; 520(7549): 612.
26. Goodman WM, Spruill SE, Komaroff E. A proposed hybrid effect size plus p-value criterion: empirical evidence supporting its use. The American Statistician 2019; 73(sup1): 168–185.

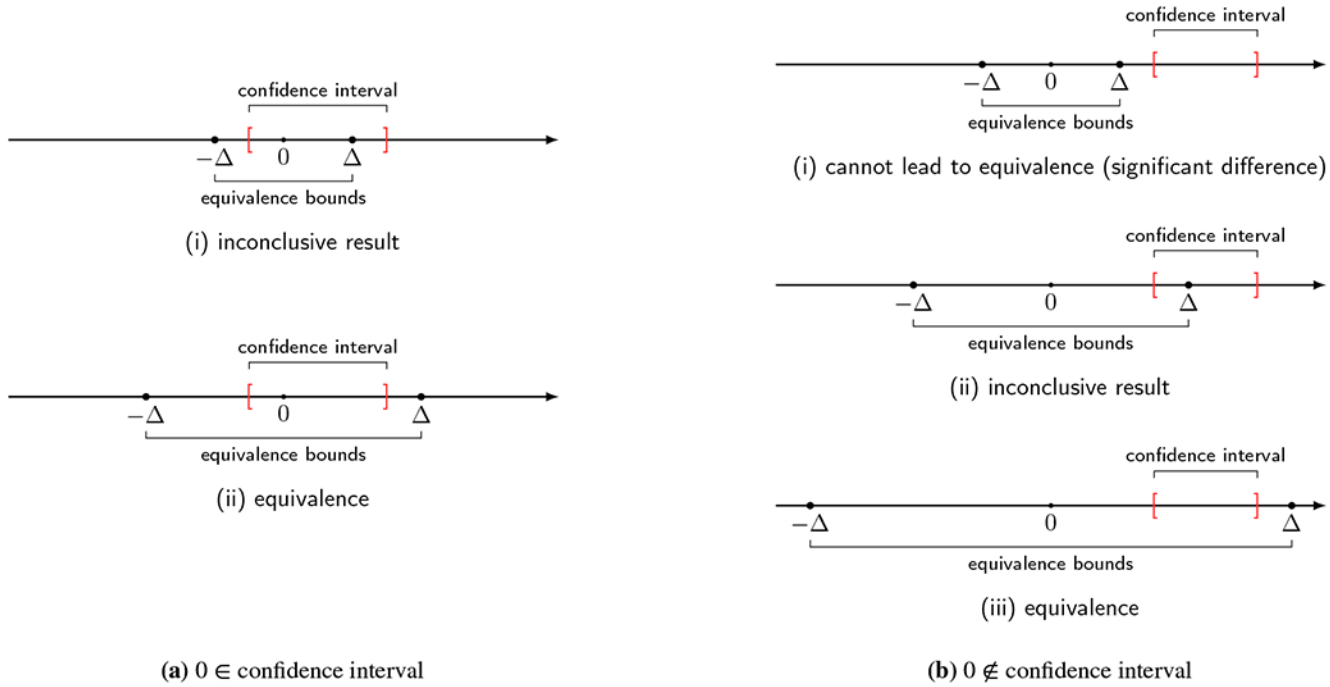
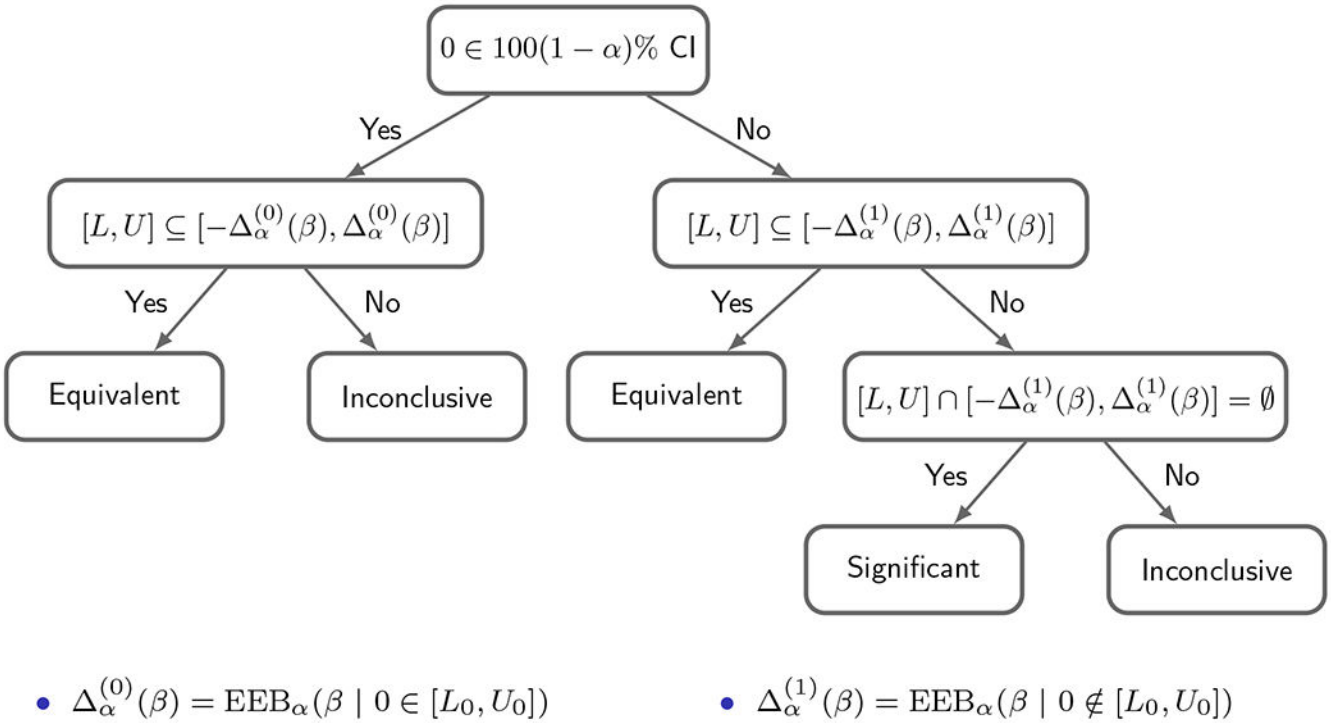


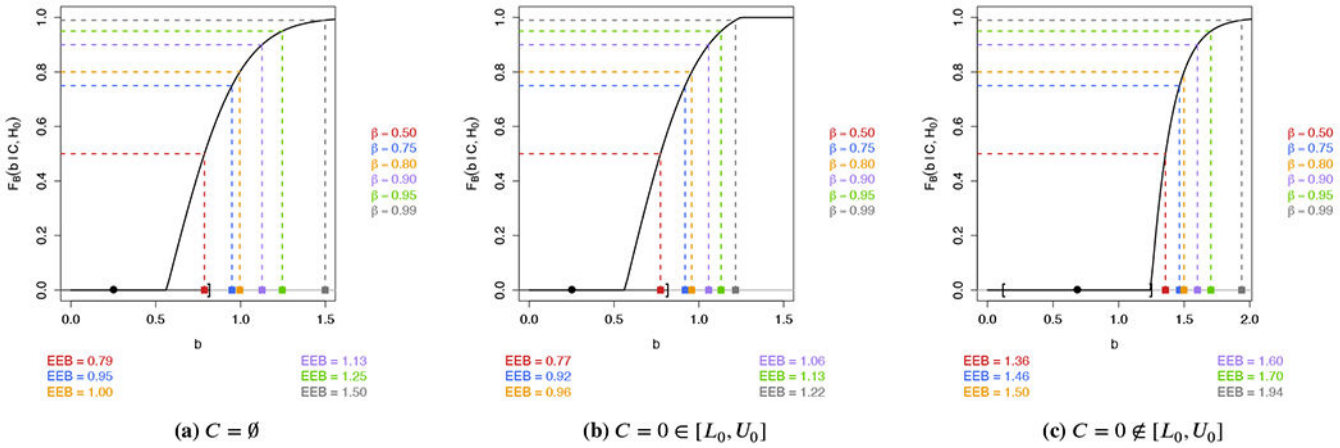
FIGURE 1.

Conclusion of an equivalence test with different prespecified equivalence bounds when (a) the confidence interval covers zero and (b) the confidence interval does not cover zero. Two possible conclusions under scenario (a): (i) inconclusive result and (ii) equivalence. Three possible conclusions under scenario (b): (i) significant difference, (ii) inconclusive result, and (iii) equivalence.



- $\Delta_\alpha^{(0)}(\beta) = \mathrm{EEB}_\alpha(\beta \mid 0 \in [L_0, U_0])$
- $\Delta_\alpha^{(1)}(\beta) = \mathrm{EEB}_\alpha(\beta \mid 0 \notin [L_0, U_0])$

FIGURE 2.
A two-stage testing procedure comparing two means.



(a) $C = \emptyset$ **(b)** $C = 0 \in [L_0, U_0]$ **(c)** $C = 0 \notin [L_0, U_0]$

FIGURE 3.

Marginal and conditional distribution of the B-value and the corresponding EEB at various β levels in the example.

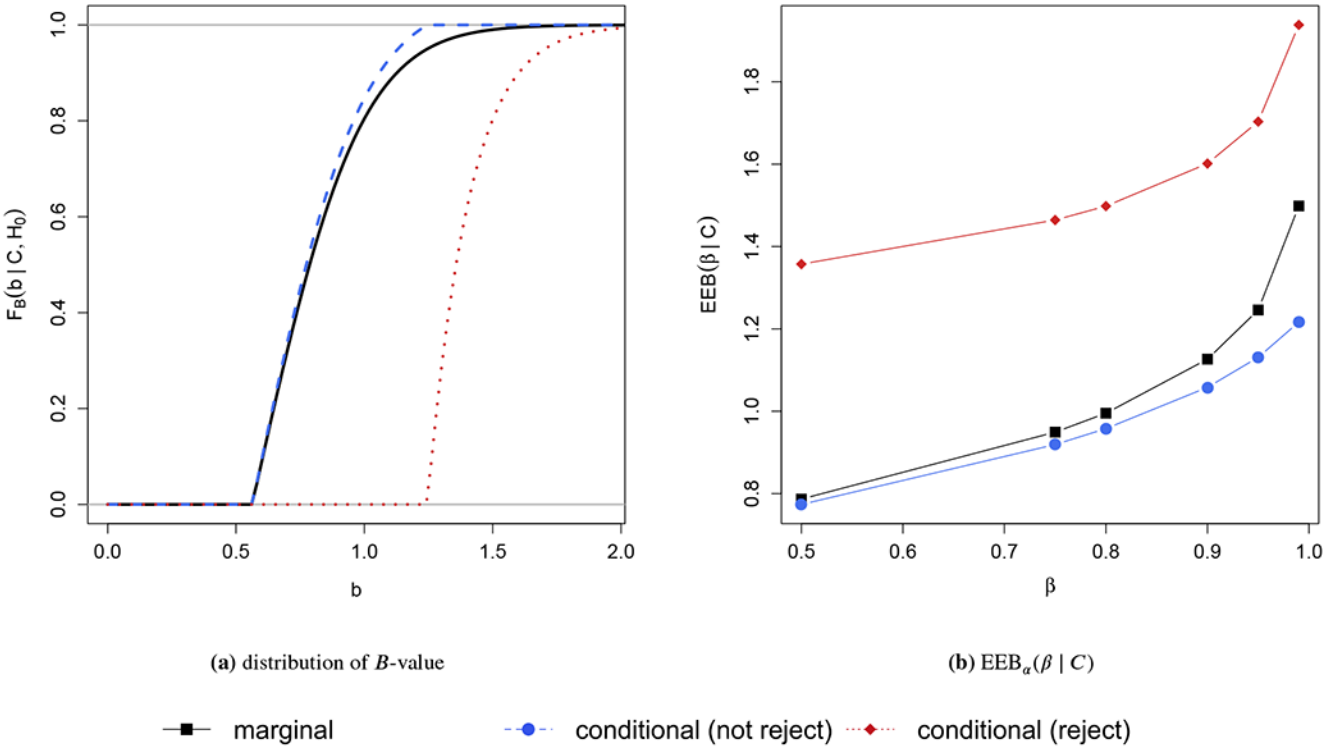


FIGURE 4.

(a) The marginal and conditional distribution of the B-value and (b) the marginal and conditional EEB at various β levels in the example.

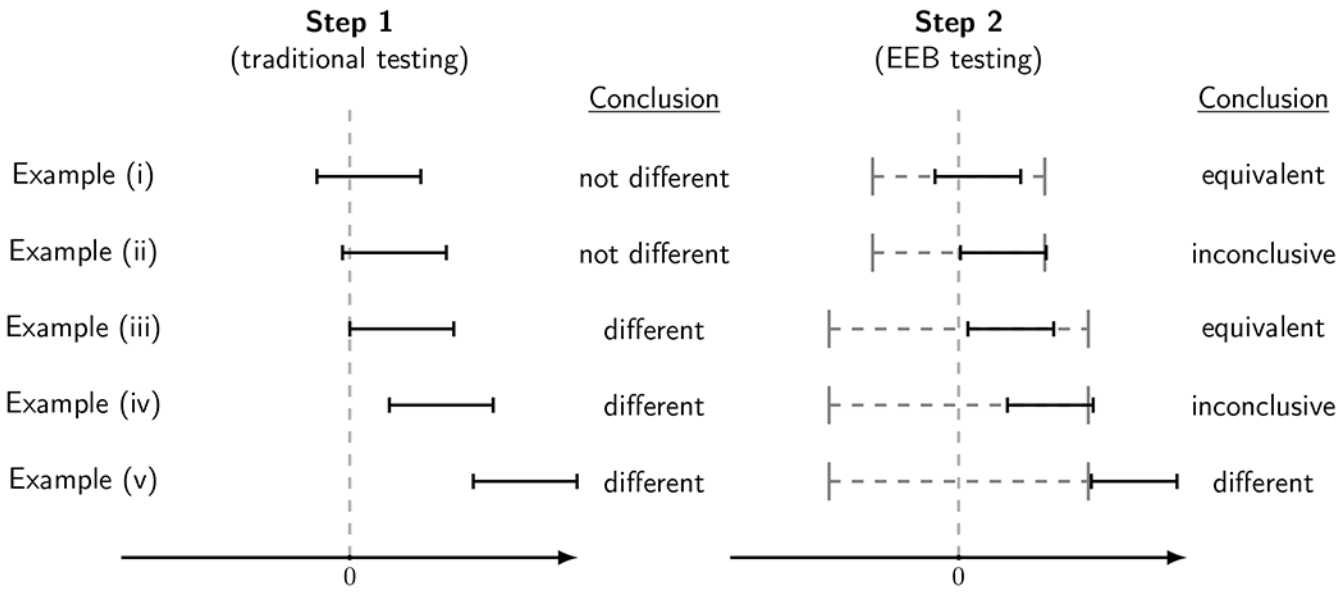
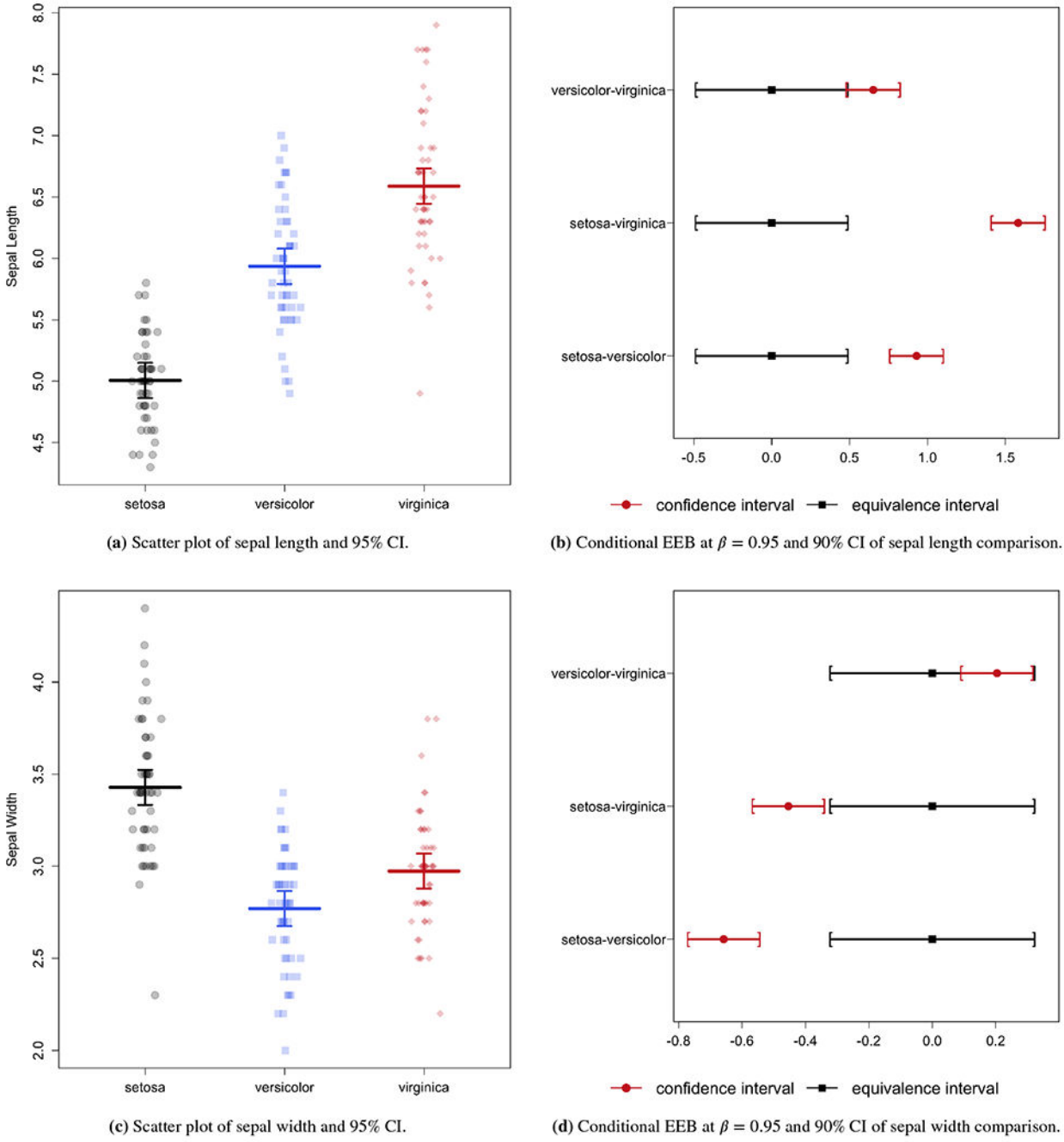


FIGURE 5.

Implementation of the two-stage testing procedure on the example data with $\alpha = 0.05$ and $\beta = 0.95$. The black solid lines in Step 1 are the 95% confidence intervals, the black solid lines in Step 2 are the 90% confidence intervals, and the gray dashed lines are the equivalence intervals with margin level at $\mathrm{EEB}_{\alpha}(\beta \mid C)$.



(a) Scatter plot of sepal length and 95% CI.

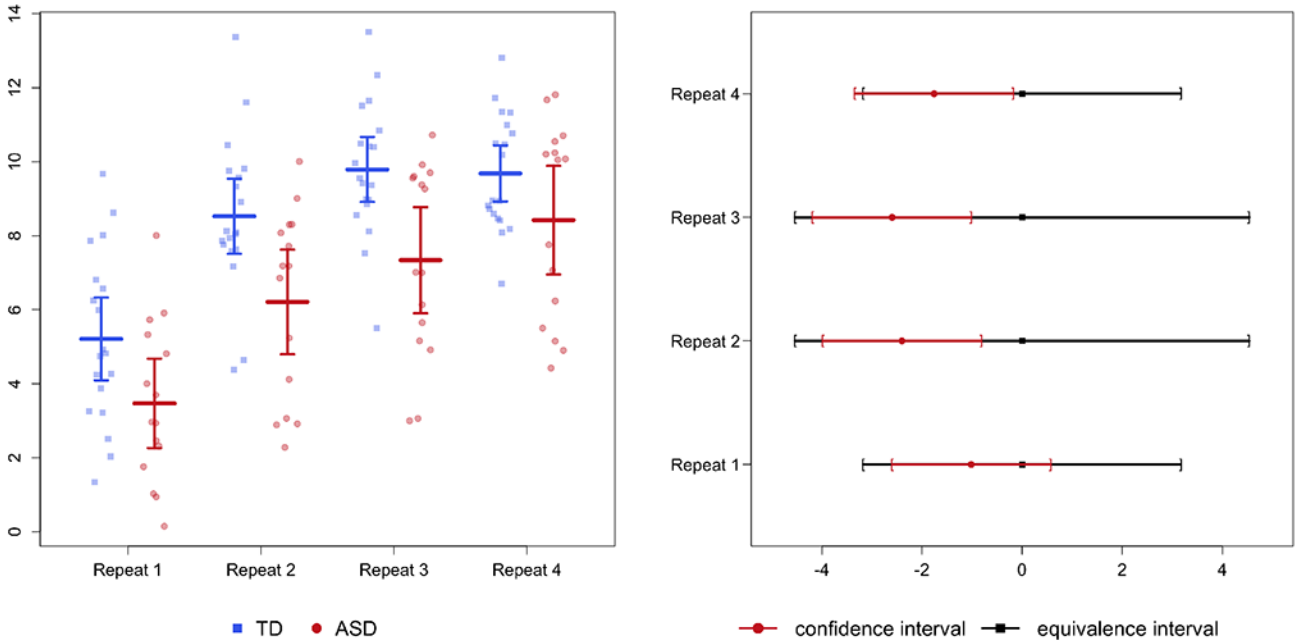
(b) Conditional EEB at $\beta = 0.95$ and 90% CI of sepal length comparison.

(c) Scatter plot of sepal width and 95% CI.

(d) Conditional EEB at $\beta = 0.95$ and 90% CI of sepal width comparison.

FIGURE 6.

(a)&(c) Data scatter plot and the 95% confidence interval (CI) of each species. (b)&(d) The conditional empirical equivalence interval with $\beta = 0.95$ and the 90% CI of each pair-wise comparison.



(a) Data distribution and 95% CI after adjusting for age, sex, and motor coordination.

(b) Conditional EEB at $\beta = 0.95$ and 90% CI.

FIGURE 7.

(a) Scatter plot and 95% confidence interval (CI) of the performance data after adjusting for age, sex, and motor coordination. (b) The conditional empirical equivalence interval with $\beta = 0.95$ and the 90% CI of the comparison at each repetition.

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 1

Statistics in the two-sample t-test using the simulated data under two scenarios: $0 \in [L_0, U_0]$ and $0 \notin [L_0, U_0]$. Five examples are considered such that the conclusion in the second step differs.

Example	$\hat{\delta}$	S	t-value	p-value	$[L_0, U_0]$	$[L, U]$
(i)	0.252	0.325	0.775	0.448	[−0.431, 0.934]	[−0.311, 0.815]
(ii)	0.587	0.325	1.807	0.087	[−0.095, 1.269]	[0.024, 1.150]
(iii)	0.685	0.325	2.108	0.049	[0.002, 1.367]	[0.121, 1.248]
(iv)	1.203	0.325	3.704	0.002	[0.521, 1.885]	[0.640, 1.766]
(v)	2.305	0.325	7.097	< 0.001	[1.623, 2.987]	[1.742, 2.868]

TABLE 2

The proportion of concluding equivalence in Step 2 testing using either the marginal empirical equivalence bound (EEB) or the conditional EEB. The proportion is reported for all 10000 simulations as well as based on the result of Step 1 testing.

	Overall	Result of Step 1 testing	
		$0 \in [L_0, U_0]$	$0 \notin [L_0, U_0]$
Marginal EEB	0.952	1.000	0.000
Conditional EEB	0.952	0.952	0.956

TABLE 3

Statistics of the comparison between iris species. The results are derived from an ANOVA model. Tukey's *p* are the *p*-values from Tukey's test.

Outcome	Comparison	Estimate (SE)	*t*-statistic	*p*-value	Tukey's *p*	95% CI	90% CI
Sepal length	setosa–versicolor	0.930 (0.103)	9.033	< 0.001	< 0.001	[0.726, 1.133]	[0.759, 1.100]
	setosa–virginica	1.582 (0.103)	15.365	< 0.001	< 0.001	[1.378, 1.785]	[1.411, 1.752]
	versicolor–virginica	0.652 (0.103)	6.333	< 0.001	< 0.001	[0.448, 0.855]	[0.481, 0.822]
Sepal width	setosa–versicolor	−0.658 (0.068)	−9.685	< 0.001	< 0.001	[−0.792, −0.524]	[−0.770, −0.545]
	setosa–virginica	−0.454 (0.068)	−6.683	< 0.001	< 0.001	[−0.588, −0.320]	[−0.566, −0.341]
	versicolor–virginica	0.204 (0.068)	3.003	< 0.001	0.009	[0.070, 0.338]	[0.091, 0.316]

TABLE 4

Statistics of the comparison between ASD and TD in the performance gesture learning for four repetitions. The results are derived from a mixed effects model with group, repetition, their interaction, as well as age, sex, and motor coordination (quantified via PANESS), as the covariates. Tukey's *p* are the *p*-values from Tukey's test.

Repetition	Estimate (SE)	*t*-statistic	*p*-value	Tukey's *p*	95% CI	90% CI
1	−1.018 (0.952)	−1.069	0.287	0.528	[−2.908, 0.872]	[−2.600, 0.563]
2	−2.401 (0.952)	−2.521	0.013	0.031	[−4.291, −0.411]	[−3.982, −0.819]
3	−2.597 (0.952)	−2.728	0.007	0.017	[−4.487, −0.707]	[−4.179, −1.016]
4	−1.759 (0.952)	−1.847	0.068	0.147	[−3.649, 0.131]	[−3.340, −0.177]