ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Post-selection Inference in Multiverse Analysis (PIMA): An Inferential Framework Based on the Sign Flipping Score Test

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., et al. (2024). Post-selection Inference in Multiverse Analysis (PIMA): An Inferential Framework Based on the Sign Flipping Score Test. PSYCHOMETRIKA, 89(2 (June)), 542-568 [10.1007/s11336-024-09973-6].

Availability:
This version is available at: https://hdl.handle.net/11585/969196 since: 2024-05-09

Published:

DOI: http://doi.org/10.1007/s11336-024-09973-6

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

06 May 2026

POST-SELECTION INFERENCE IN MULTIVERSE ANALYSIS (PIMA): AN INFERENTIAL FRAMEWORK BASED ON THE SIGN FLIPPING SCORE TEST

Paolo Girardi[1], Anna Vesely[2], Daniël Lakens[3], Gianmarco Altoè[4], Massimiliano Pastore[4], Antonio Calcagnì[4,5], Livio Finos[6]

[1]department of environmental sciences, informatics and statistics, ca' foscari university of venice, venezia, italy

[2]department of statistical sciences, university of bologna, bologna, italy

[3]department of industrial engineering and innovation sciences, eindhoven university of technology, eindhoven, netherlands

[4]department of developmental psychology and socialisation, university of padova, italy

[5]gncs research group, national institute of advanced mathematics (indam), rome, italy

[6]department of statistical sciences, university of padova, padova, italy

Contact Info

Paolo Girardi: paolo.girardi@unive.it; Anna Vesely: anna.vesely2@unibo.it; Daniël Lakens: D.Lakens@tue.nl; Gianmarco Altoè: gianmarco.altoe@unipd.it; Massimiliano Pastore: massimiliano.pastore@unipd.it; Antonio Calcagnì: antonio.calcagni@unipd.it; Livio Finos: livio.finos@unipd.it.

Founding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

All R code and data associated with the real data application are available at `https://osf.io/3ebw9/`, while further analyses can be developed trough the dedicated package Jointest (Finos, 2022) available at `https://github.com/livioivil/jointest`

Correspondence should be sent to

Paolo Girardi
Address: Via Torino 155, 30172 Venezia-Mestre (VE), Italy
E-Mail: paolo.girardi@unive.it

POST-SELECTION INFERENCE IN MULTIVERSE ANALYSIS (PIMA): AN INFERENTIAL FRAMEWORK BASED ON THE SIGN FLIPPING SCORE TEST

Abstract

When analyzing data, researchers make some choices that are either arbitrary, based on subjective beliefs about the data-generating process, or for which equally justifiable alternative choices could have been made. This wide range of data-analytic choices can be abused and has been one of the underlying causes of the replication crisis in several fields. Recently, the introduction of multiverse analysis provides researchers with a method to evaluate the stability of the results across reasonable choices that could be made when analyzing data. Multiverse analysis is confined to a descriptive role, lacking a proper and comprehensive inferential procedure. Recently, specification curve analysis adds an inferential procedure to multiverse analysis, but this approach is limited to simple cases related to the linear model, and only allows researchers to infer whether at least one specification rejects the null hypothesis, but not which specifications should be selected. In this paper we present a Post-selection Inference approach to Multiverse Analysis (PIMA) which is a flexible and general inferential approach that considers for all possible models, i.e., the multiverse of reasonable analyses. The approach allows for a wide range of data specifications (i.e. preprocessing) and any generalized linear model; it allows testing the null hypothesis that a given predictor is not associated with the outcome, by combining information from all reasonable models of multiverse analysis, and provides strong control of the family-wise error rate allowing researchers to claim that the null hypothesis can be rejected for any specification that shows a significant effect. The inferential proposal is based on a conditional resampling procedure. We formally prove that the Type I error rate is controlled, and compute the statistical power of the test through a simulation study. Finally, we apply the PIMA procedure to the analysis of a real dataset on the

self-reported hesitancy for the COronaVIrus Disease 2019 (COVID-19) vaccine before and after the 2020 lockdown in Italy. We conclude with practical recommendations to be considered when implementing the proposed procedure.

Key words: multiverse analysis, flipping score, statistical inference, testing, reproducibility, replicability

1. Introduction

Real-world data analysis often involves many defensible choices at each step of the analysis, such as how to combine and transform measurements, how to deal with missing data and outliers, and even how to choose a statistical model. In general, there is not a single defensible choice for every decision researchers must make, and there are many defensible options for each step of the data analysis (Gelman and Loken, 2014). As a result, raw data do not uniquely yield a single dataset for analysis. Instead, researchers are faced with a set of processed datasets, each determined by a unique combination of choices – a multiverse of datasets. Since analyses performed on each dataset may yield different results, the data multiverse directly implies a multiverse of statistical results. In recent years, concerns have been raised about how researchers can exploit this flexibility in data analysis to increase the likelihood of observing a statistically significant result. Researchers may engage in such questionable research practices due to editorial practices that prioritize the publication of statistically significant results or the selection of findings that confirm the belief of the same authors (Begg and Berlin, 1988; Dwan et al., 2008; Fanelli, 2012). When researchers select and report the results of a subset of all possible analyses that produce significant results (Sterling, 1959; Greenwald, 1975; Simmons et al., 2011; Brodeur et al., 2016), they dramatically increase the actual false-positive rates despite their nominal endorsement of a low Type I error rate (e.g., 5%).

Two solutions have been proposed to address the issue of p-hacking. The first solution requires researchers to specify their statistical analysis plan before examining raw data. Such preregistered studies control the Type I error rate by reducing flexibility during the data analysis. Preregistration is easily implemented for replication studies, where researchers specify that they will perform the same analysis as was performed in an earlier study. For more novel studies, preregistration can be challenging because researchers may not have enough knowledge to anticipate all the possible decisions that need to be made when analyzing the data. The second solution recognizes that it is often not feasible to specify a single analysis before collecting the data and instead advocates for transparently reporting all possible analyses that can be conducted. Steegen et al. (2016) introduced multiverse analysis, which aims to use all reasonable

options for data processing to construct a multiverse of datasets, and then separately perform the same analysis of interest on each of these datasets. The main tool used to interpret the output of a multiverse analysis is a histogram of p-values, which summarizes all the p-values obtained for a given effect. Researchers then typically discuss the results in terms of the proportion of significant p-values. This procedure not only provides a detailed picture of the robustness or fragility of the results in different processing choices, but also allows researchers to explore the key choices that are most consequential in the fluctuation of their results. Multiverse analysis represents a valuable step towards transparent science. The method has gained popularity since its development and has been applied in various experimental contexts, including cognitive development, risk perception (Mirman et al., 2021), assessment of parental behavior (Modecki et al., 2020), and memory tasks (Wessel et al., 2020). Although some applications are limited to exploratory purposes, aiming to define brief guidelines for conducting a multiverse analysis (Dragicevic et al., 2019; Liu et al., 2020), other studies use this method as a robustness assessment for mediation analysis (Rijnhart et al., 2021) or an exhaustive modeling approach (Frey et al., 2021). This research approach permits to exhibit the stability and robustness of findings, not only across different exclusion criteria or modifications of variables, but also across different decisions made during all phases of data analysis. This feature can be particularly interesting and appealing from the perspective of the replicability crisis in quantitative psychology (Open Science Collaboration, 2015), and in enhancing the transparency and credibility of scientific results (Nosek and Lakens, 2014). Multiverse analysis can therefore be extended beyond the pre-processing stage to include the methods used for the analysis (the "multiverse of methods") (Harder, 2020). The explicit flexibility in multiverse analysis is not to be condemned as it reflects an effort to transparently describe the uncertainty about the best analysis strategy. However, if, on the one hand, the exploration of multiple analytical choices in data analysis must be advocated, on the other hand it is challenging to draw reliable inferences from such a large number of statistical analyses. Although most researchers have interpreted the results derived from multiverse analysis descriptively, while doing so, it is extremely tempting to make claims about analyses that yield statistically significant results, and not to make claims about non-significant results. However, a selective focus on a subset of statistically significant results

once again introduces the problem of selective inference (Benjamini, 2020), and can potentially inflate the rate at which claims about effects are false positives.

Currently, the only method that allows researchers to make formal inferences in multiverse analysis is specification curve analysis (Simonsohn et al., 2020). Analogously to multiverse analysis, it requires researchers to consider the entire set of reasonable combinations of data-analytic decisions, called specifications; subsequently, these specifications are used jointly to derive a test for the null hypothesis of interest. If the null hypothesis is rejected, researchers can claim with a certain maximum error rate (e.g., 5%) that there exists at least one specification in which the null hypothesis is false. In the most general case of non-experimental data, the inferential support is based on bootstrapping techniques and is valid only in linear regression models (LMs), without the possibility of a general extension to other distributions for the dependent variable that are usually included in generalized linear models (GLMs). More importantly, this methodology lacks a formal description of the statistical properties of the test, allows testing only a single hypothesis, and does not address the problem of controlling multiplicity when testing different hypotheses. A more formal study of the method's performance is provided in Sections 3 and 4. Because researchers are often interested in models that are more complex than LMs, want to explore several different processing steps, and possibly wish to investigate more null hypotheses together, it would be beneficial if more advanced analysis methods for multiverse analysis were developed. Such more advanced methods would allow, for example, psychometricians to identify a set of predictors that are associated with a particular outcome, or allow neuroscientists to identify brain regions activated by a stimulus. In summary, the multiverse analysis framework allows researchers to manage degrees of freedom in the data analysis, but the literature still lacks a formal inferential approach that allows researchers to derive reliable inferences about (sets of) specific analyses included in multiverse analysis. In this paper, we define the Post-selection Inference approach to Multiverse Analysis (PIMA) which is a flexible and general inference approach for multiverse analysis that accounts for all possible models, i.e., the multiverse of reasonable analyses. In the framework of GLMs, we consider the null hypothesis that a given predictor of interest is not associated with the outcome, i.e., that the corresponding coefficient is zero. Furthermore, we assume that researchers consider all reasonable

models obtained by different choices of data processing. We provide a resampling-based procedure based on the sign-flip score test of Hemerik et al. (2020) and De Santis et al. (2022) that allows researchers to test the null hypothesis by combining information from all reasonable models, and show that this framework allows inference about the coefficient of interest on three different levels of complexity. First, considering the predictor of interest, we compute a *global p-value* considering all models, so that researchers can state whether the coefficient is non-null in at least one of the models in the multiverse analysis. Second, we compute individual *adjusted p-values* for each model and thus obtain the set of models where the coefficient is non-null. Because PIMA accounts for multiplicity, researchers are free to choose the preferred model post-hoc, after trying all models and seeing the results. In other words, the procedure allows selective inference, but unlike p-hacking, researchers can select statistically significant analyses from the multiverse while controlling the Type I error rate. Finally, we define a third inference strategy for multiverse analysis in which researchers provide a lower confidence bound for the *true discovery proportion* (TDP), i.e., the proportion of models with a non-null coefficient. In this analysis, researchers cannot individually identify statistically significant models in the multiverse, but in some cases it may be more powerful to report the true discovery proportion than individual p-values. Finally, we argue that the method can be easily extended to the case of multiple hypotheses on different coefficients. The resulting procedure is general, flexible, and powerful, and can be applied to many different contexts. It is valid as long as all the considered models are reasonable and specified in advance, before carrying out the analysis. The structure of the paper is as follows. In Section 2 we define the framework and construct the desired resampling-based test. Subsequently, in Section 3, we use the test to make inference in the multiverse framework. We then study the properties of the PIMA method and we apply it to real data in Sections 4 and 5, respectively. We conclude with Section 6 that contains a short remark on the main results, with some hints on still open issues in multiverse analysis and practical recommendations for the PIMA methodology. All the analyses and simulations were implemented using the statistical software R (R Core Team, 2021). All R code and data associated with the real data application are available at `https://osf.io/3ebw9/`, while further analyses can be developed through the dedicated package Jointest (Finos, 2022) available at `https://github.com/livioivil/jointest`.

2. The sign-flip score test

In the context of multiverse analysis, there is no a single pre-specified model, while we are interested in testing the effect of a given predictor on a response variable in the multiverse of possible models. In order to test the global null hypothesis that the predictor has no effect in any of the models considered, one needs to define a proper test statistic and its distribution under the null hypothesis. Finding a solution within the parametric framework represents a formidable challenge, due to the inherent dependence among the univariate test statistics, which in most cases is very high and usually non-linear. The resampling-based approach usually provides a solution to this multivariate challenge. We will rely on the sign-flip score test of Hemerik et al. (2020) and De Santis et al. (2022) to define an asymptotically exact test for the global null hypothesis of interest. In this section, we specify the structure of the models and introduce the sign-flip score test for a single model specification. In the next section, we will give a natural extension to the multivariate framework. Finally, we will show how to employ the procedure within the closed testing framework (Marcus et al., 1976) to make additional inferences on the models.

2.1. Model specification

We consider the framework of GLMs. Let $Y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ be n independent observations of a variable of interest, which is assumed to belong to the exponential dispersion family distribution with density of the form

$$h(y_i, \theta_i, \phi_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\} \qquad (i = 1, \dots, n),$$

where θ_i and ϕ_i are the canonical and the dispersion parameter, respectively. According to the usual literature of GLMs (Agresti, 2015), the mean and variance functions are

$$\mu_i = E[y_i] = b'(\theta_i), \qquad v(\mu_i) = b''(\theta) = \frac{\mathrm{var}(y_i)}{a(\phi_i)}.$$

We suppose that the mean of Y depends on an observed predictor of interest $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$ and m other observed predictors $Z = (z_1, \dots, z_n)^\top \in \mathbb{R}^{n \times m}$ through a non-linear relation

$$g(\mu_i) = \eta_i = x_i\beta + z_i\gamma$$

where $g(\cdot)$ denotes the link function, $\beta \in \mathbb{R}$ is a parameter of interest, and $\gamma \in \mathbb{R}^m$ is a vector of nuisance parameters.

Finally we define the following $n \times n$ matrices , that will be used in the next sections:

$$D = \text{diag}\{d_i\} = \text{diag}\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\}$$
$$V = \text{diag}\{v_i\} = \text{diag}\{\text{var}(y_i)\}$$
$$W = DV^{-1}D.$$

2.2. Hypothesis testing for an individual model via sign-flip score test

Given a model specified as in the previous section, we are interested in testing the null hypothesis $\mathcal{H} : \beta = 0$ that the predictor X does not influence the response Y with significance level $\alpha \in [0, 1)$. Here γ is estimated by $\hat{\gamma}$, and is therefore a vector of nuisance parameters. We consider the hypothesis $\beta = 0$ for simplicity of exposition, however the sign-flip approach can be extended to the more general case $\beta = \beta_0$.

Relying on the work of Hemerik et al. (2020), De Santis et al. (2022) provide the sign-flip score test, a robust and asymptotically exact test for $\mathcal{H}$ that uses B random sign-flipping transformations. Even though larger values of B tend to give more power, to have non-zero power it is sufficient to take $B \geq 1/\alpha$. Hence, consider the $n \times n$ diagonal matrices $F^b = \text{diag}\{f_i^b\}$, with $b = 1, \ldots, B$. The first is fixed as the identity $F^1 = I$, and the diagonal elements of the others are independently and uniformly drawn from $\{-1, 1\}$. Each matrix F^b defines a flipped effective score

$$S^b = n^{-1/2} X^\top W^{1/2}(I - Q)V^{-1/2}F^b(Y - \hat{\mu}) \tag{1}$$

where

$$Q = W^{1/2}Z(Z^\top WZ)^{-1}Z^\top W^{1/2}$$

is a particular hat matrix, symmetric and idempotent, and $\hat{\mu}$ is a $\sqrt{n}$-consistent estimate of the true value μ^* computed under $\mathcal{H}$. In practical applications, if the matrices D and V, and thus W, are unknown, they can be replace by $\sqrt{n}$-consistent estimates.

This effective score may be written as a sum of individual contributions with flipped signs, as follows,

$$S^b = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_i^b \, \nu_i, \qquad \nu_i = \left(x_i - X^\top W Z (Z^\top W Z)^{-1} z_i \right) \frac{(y_i - \hat{\mu}_i) d_i}{v_i}. \tag{2}$$

Here ν_i is the contribution of the i-th observation to the effective score. The definition and properties of the contributions ν_i are explored in Hemerik et al. (2020) and De Santis et al. (2022), where they are denoted as $\nu^*_{\hat{\gamma},i}$ and $\tilde{\nu}^*_{i,\beta}$, respectively.

An assumption is needed about the effective score computed when the true value γ^* of the nuisance γ, and so the true value μ^* of μ, are known. This quantity may be written analogously to (1) and (2), as

$$S^{*b} = n^{-1/2} X^\top W^{1/2} (I - Q) V^{-1/2} F^b (Y - \mu^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_i^b \, \nu_i^*. \tag{3}$$

In this case, the contributions ν_i^* are independent if D and V are known, and asymptotically independent otherwise (Hemerik et al., 2020). The required assumption is a Lindeberg's condition that ensures that the contribution of each ν_i^* to the variance of S^{b*} is arbitrarily small as n grows. This can be formulated as follows.

Assumption 1. As $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \operatorname{var}(\nu_i^*) \longrightarrow c$$

for some constant $c > 0$. Moreover, for any $\varepsilon > 0$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left(\nu_i^{*2} \cdot \mathbf{1} \left\{ \frac{|\nu_i^*|}{\sqrt{n}} > \varepsilon \right\} \right) \longrightarrow 0$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

Given this assumption, the sign-flip score test of De Santis et al. (2022) relies on the standardized flipped scores, obtained by each effective score (1) by its standard deviation:

$$\tilde{S}^b = S^b \operatorname{var}(S^b \mid F^b)^{-1/2} \tag{4}$$

where

$$\operatorname{var}(S^b \mid F^b) = n^{-1} X^\top W^{1/2} (I - Q) F^b (I - Q) F^b (I - Q) W^{1/2} X + o_P(1).$$

The test is defined from the absolute values of the standardized scores, comparing the observed value $|\tilde{S}^1|$ with a critical value obtained from permutations. The latter is $|\tilde{S}|^{\lceil(1-\alpha)B\rceil}$, where $|\tilde{S}|^{(1)} \leq \ldots \leq |\tilde{S}|^{(B)}$ are all the sorted values and $\lceil\cdot\rceil$ denotes the ceiling function.

Theorem 1. (De Santis et al., 2022) Under Assumption 1, the test that rejects $\mathcal{H}$ when $|\tilde{S}^1| > |\tilde{S}|^{\lceil(1-\alpha)B\rceil}$ is an α-level test, asymptotically as $n \to \infty$.

The test of Theorem 1 is exact in the particular case of LMs, and second-moment exact in GLMs. The second-moment exactness means that under $\mathcal{H}$ the test statistics $\tilde{S}^b$ do not necessarily have the same distribution, but share the same mean and variance, independently of the sign flip; this provides exact control of the Type I error rate, for practical purposes, even for finite sample size. The only requirement is Assumption 1, that states that the variance of the score (3) is not dominated by any particular contribution. Furthermore, the test is robust to some model misspecifications, as long as the mean μ and the link g are correctly specified. In particular, under minimal assumptions, the test is still asymptotically exact for any generic misspecification of the variance V (De Santis et al., 2022).

2.3. Intuition behind the sign-flip score test

Although the formal definition of the sign-flip score approach may seem difficult to grasp, its meaning is quite intuitive. For the sake of clarity, we consider a simple example using a GLM with gaussian error and identity link, that can be easily reconducted to a multiple linear model. In this case, we have $W = D = I$ and $V = \sigma^2 I$, where σ^2 is the variance shared by every observation. From (2), the observed and flipped effective scores can be written as

$$S^1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nu_i, \qquad S^b = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_i^b \, \nu_i \qquad (b = 2, \ldots, B)$$

where

$$\nu_i = \frac{1}{\sigma^2}(x_i - \hat{x}_i)(y_i - \hat{y}_i), \qquad \hat{x}_i = X^\top Z (Z^\top Z)^{-1} z_i, \qquad \hat{y}_i = \hat{\mu}_i = Y^\top Z (Z^\top Z)^{-1} z_i. \tag{5}$$

In this perspective, the score can be interpreted as the sum of weighted residuals of $y_i - \hat{y}_i$, where the weights are the residuals $x_i - \hat{x}_i$. A further interpretation is that the score is the sum of n

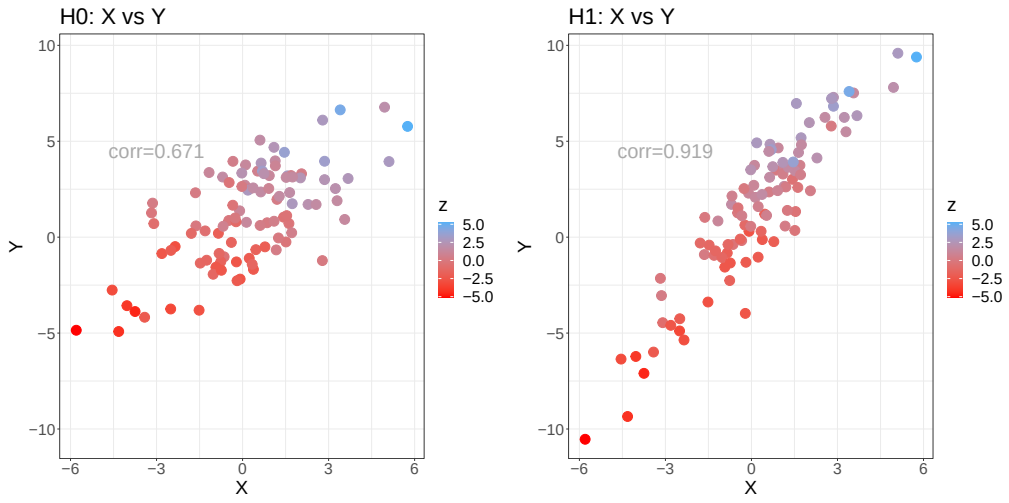

FIGURE 1.

Simulated dataset under the scenario H_0 (left) and H_1 (right).

contributions, and these contributions are the residuals of y_i predicted by z_i multiplied by the residuals of x_i predicted by z_i. In this sense, the score extends the covariance by moving from the empirical mean (i.e., a model with the intercept only) to a full linear model.

To see things in practice, consider the following linear regression model

$$Y = 1 + \beta X + \gamma Z + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_n(0, I)$$

and suppose we are interested in testing $\mathcal{H} : \beta = 0$. The predictors X and Z are generated from a multivariate normal with unit variance and covariance 0.80. We create two scenarios, sharing the same X and Z, but with different response variable Y: the first scenario is generated under the null hypothesis $\mathcal{H}$ ($\beta = 0$, $\gamma = 1$), while the second is generated under the alternative ($\beta = 1$, $\gamma = 1$). For each simulation, we generate $n = 100$ observations. We name the resulting datasets H_0 and H_1, respectively.

Examples of scatter plots between Y and X considering also Z by color are given in Figure 1. In both scenarios we see a positive correlation between X and Y. From the color of the dots one can appreciate the positive dependence of Z – both – with X and Y; that is, more bluish dots correspond to higher values of Z and these appear where X and Y have higher values too (upper right corner). Testing for the null hypothesis $\mathcal{H} : \beta = 0$, however, corresponds to testing the

partial correlation between X and Y, net of the effect of Z. This partial correlation can be visually evaluated with a scatter plot of the residuals that form the n addends of the observed score S^1 given in (5). These are shown in the two upper plots of Figure 2 for both the datasets H_0 (upper left) and H_1 (upper right). In these scatter plots the coordinates of each point are the values $x_i - \hat{x}_i$ and $y_i - \hat{y}_i$, and the observed score S^1 is obtained from the sum of the product of these coordinates $(x_i - \hat{x}_i)(y_i - \hat{y}_i)$. After removing the effect of Z from X and Y, the scatter plot for H_0 shows no relationship between the two variables, while this is still present for H_1.

The distribution of the effective score under the null hypothesis $\mathcal{H}$ is obtained by computing a large number of flipped scores S^b. Each flipped score is determined by randomly flipping the signs of the score contributions ν_i, and thus of the residuals $(y_i - \hat{y}_i)$. The effect of these sign flips is visible in the scatter plots at the bottom of the Figure 2. The positive (partial) correlation of the H_1 dataset (top right) is destroyed by random flips and is now approximately zero (bottom right). For the H_0 dataset a random flip maintains the observed correlation around zero.

There are two further delicate details that may provide an additional value to the flip-scores approach: 1) the need for sign flips instead of permutations; 2) the need for the standardization step. One may notice that, in the example proposed here, one may permute the residuals instead of flipping the signs. However, this is only true in the context of homoscedasticity, but would not be a valid option in the more general case of GLMs. Although the intuition provided here holds also for the GLM, one has to bear in mind that the zero-centered contributions ν_i (2) should be such that $\text{var}(\nu_i) = \text{var}(-\nu_i)$, while this would not hold when permuting the residuals $(y_i - \hat{y}_i)$.

The second relevant detail is the standardization step of De Santis et al. (2022), introduced in (4). Due to the fact that the nuisance parameters γ are unknown and must be estimated, the residuals $(y_i - \hat{y}_i)$ are independent only asymptotically. As a result, asymptotically the variances of the observed and permuted scores are equal, which ensures control of the Type I error; however, this generally does not hold for finite sample sizes. The standardization step compensates for this different variability, guaranteeing exactness under the linear normal model and second-moment exactness in the more general GLM setting.

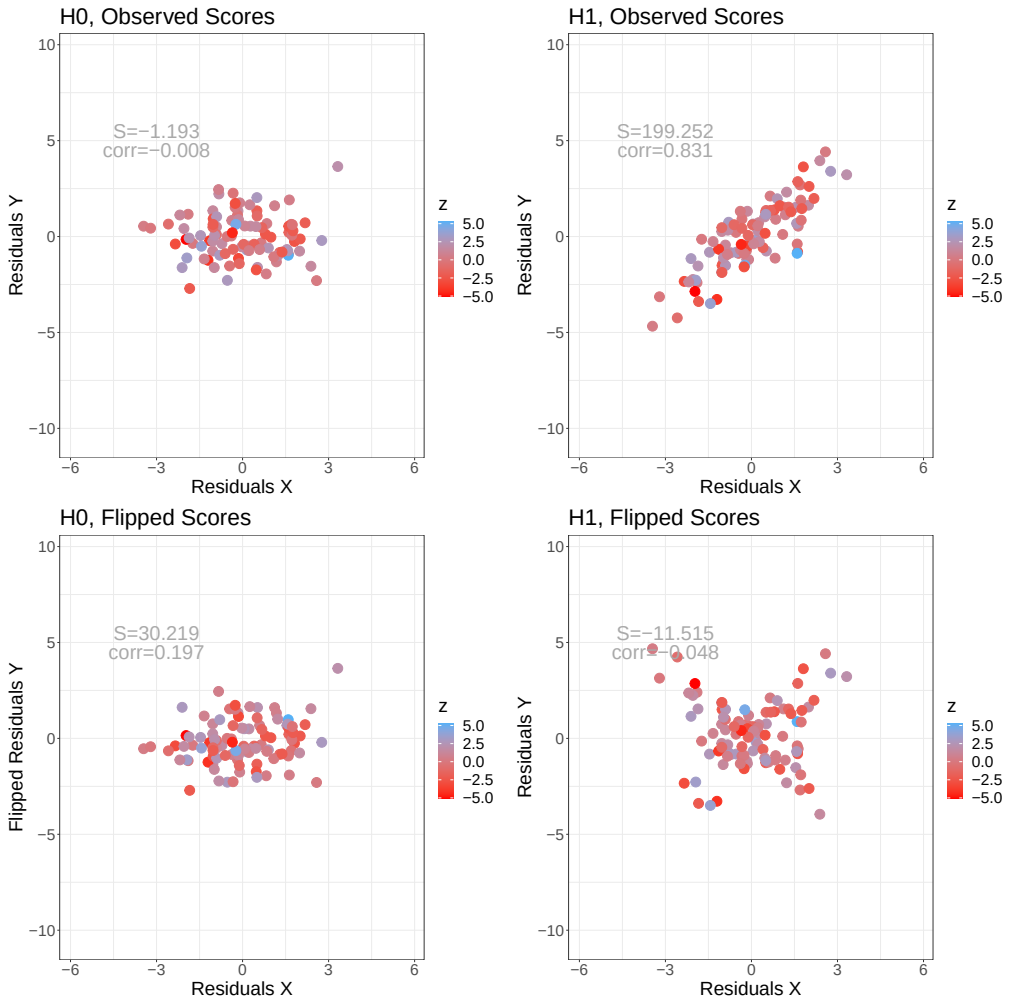

FIGURE 2.

Observed (top) and flipped (bottom) distribution residuals of Y vs. X in the datasets H_0 (left) and H_1 (right).

3. PIMA: Post-selection Inference in Multiverse Analysis

3.1. Hypothesis testing in the multiverse via combination of sign-flip score tests

In the previous section we presented an asymptotically exact test for a prefixed null hypothesis. Now we consider the framework of multiverse analysis, where we define K plausible models, given by different processing of the data. Each model $k = 1, \ldots, K$ can be characterized by different specifications of the response Y_k (e.g., by deleting outliers or removing leverage points), of the predictors X_k and Z_k (e.g., by combining and transforming variables), and of the link function g_k. Let β_k be the coefficient of interest in the model k, and define the null hypothesis $\mathcal{H}_k : \beta_k = 0$ analogously to the previous section. Then consider the global (i.e., multivariate) null hypothesis as the intersection of the K individual hypotheses:

$$\mathcal{H} = \bigcap_{k=1}^{K} \mathcal{H}_k \ : \ \beta_k = 0 \text{ for all } k = 1, \ldots, K.$$

This global hypothesis $\mathcal{H}$ is true when the predictor of interest has no relationship with the response in any of the K models; it is false when such a relationship exists in at least one of the models. To test $\mathcal{H}$, we will extend the test of Theorem 1 similarly to the extension given in the case of the linear model of Vesely et al. (2022).

To construct the desired global test, we first compute the flipped standardized scores (4) for all models, using the same sign-flipping transformations. Hence we obtain $\tilde{S}_1^b, \ldots, \tilde{S}_K^b$ for $b = 1, \ldots, B$. Intuitively, the n scalar contributions ν_i in (2) are now n vectors of length K, each containing the contributions of the i-th observation to each one of the K models. The same sign-flip for observation i, f_i^b, is therefore applied to the whole vector. This resampling strategy ensures that the test has an asymptotically exact control of the Type I error.

Subsequently we combine these flipped standardized scores through any function $\psi : \mathbb{R}^K \longrightarrow \mathbb{R}$ that is non-decreasing in each argument, such as the (weighted) mean and the maximum. This give us the global test statistic

$$T^b = \psi\left(|\tilde{S}_1^b|, \ldots, |\tilde{S}_K^b|\right) \qquad (b = 1, \ldots, B). \tag{6}$$

The following theorem gives a test for $\mathcal{H}$ that relies on $T^1, \ldots, T^B$.

Theorem 2. Suppose that Assumption 1 holds for all the considered models. Then the test that rejects $\mathcal{H}$ when $T^1 > T^{\lceil(1-\alpha)B\rceil}$ is an α-level test, asymptotically as $n \to \infty$.

Proof. Throughout the proof, we will denote the k-th model adding a subscript k to the corresponding quantities that vary between models. First, for simplicity of notation we consider only specifications that maintain the sample size, while we do not consider outlier deletion or leverage point removal. In this way, the response vector Y is the same across models.

Fix any $k \in \{1, \ldots, K\}$, and assume that $\mathcal{H}_k : \beta_k = 0$ is true, so that the coefficient of interest is null in the k-th model. The flipped effective scores (1) are

$$S_k^b = n^{-1/2} X_k^\top W_k^{1/2} (I - Q_k) V_k^{-1/2} F^b (Y - \hat{\mu}_k) \qquad (b = 1, \ldots, B)$$

where

$$Q_k = W_k^{1/2} Z_k (Z_k^\top W_k Z_k)^{-1} Z_k^\top W_k^{1/2}$$

and $\hat{\mu}_k$ is a $\sqrt{n}$-consistent estimate of the true value μ_k^* computed under $\mathcal{H}_k$. Consider the flipped effective scores computed when the true value γ_k^* of the nuisance γ_k, and so the true value μ_k^* of μ_k, are known as in (3),

$$S_k^{*b} = n^{-1/2} X_k^\top W_k^{1/2} (I - Q_k) V_k^{-1/2} F^b (Y - \mu_k^*) \qquad (b = 1, \ldots, B).$$

Hemerik et al. (2020) show that S_k^b and S_k^{*b} are asymptotically equivalent as $n \to \infty$ (see the proof of Theorem 2).

Subsequently, assume that the global null hypothesis $\mathcal{H}$ is true. Hence all individual hypotheses $\mathcal{H}_k$ are true, and β_k is null in all considered models. Consider the KB-dimensional vectors of effective scores

$$S = (S_1^1, \ldots, S_1^B, \ldots, S_K^1, \ldots, S_K^B)^\top$$
$$S^* = (S_1^{*1}, \ldots, S_1^{*B}, \ldots, S_K^{*1}, \ldots, S_K^{*B})^\top$$

which are asymptotically equivalent. For any couple of models $k, j \in \{1, \ldots, K\}$ and any pair of

transformations $b, c \in \{1, \dots, B\}$, we have

$$\begin{aligned}
\mathbb{E}(S_k^{*b}) &= 0 \\
\operatorname{cov}(S_k^{*b}, S_j^{*c}) &= n^{-1} X_k^\top W_k^{1/2} (I - Q_k) V_k^{-1/2} \mathbb{E}\left(F^b (Y - \mu_k^*)(Y - \mu_j^*)^\top F^c\right) V_j^{-1/2} (I - Q_j) W_j^{1/2} X_j \\
&= \begin{cases} \xi_{kj} & \text{if } b = c \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where

$$\xi_{kj} = n^{-1} X_k^\top W_k^{1/2} (I - Q_k) V_k^{-1/2} \operatorname{diag}\left((Y - \mu_k^*)(Y - \mu_j^*)^\top\right) V_j^{-1/2} (I - Q_j) W_j^{1/2} X_j.$$

Note that S^* can be written as the sum of n independent vectors. As Assumption 1 holds for all models, by the multivariate Lindeberg-Feller central limit theorem (van der Vaart, 1998)

$$S, S^* \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}_{NB}\left(\mathbf{0}, \Xi \otimes I\right)$$

where $\mathcal{N}$ denotes the multivariate normal distribution, $\otimes$ is the Kronecker product, and

$$I \in \mathbb{R}^{B \times B}, \qquad \Xi = \left(\lim_{n\to\infty} \xi_{kj}\right) \in \mathbb{R}^{K \times K}.$$

Equivalently, we can say that

$$\begin{pmatrix} S_1^1 & \dots & S_K^1 \\ \vdots & & \vdots \\ S_1^B & \dots & S_K^B \end{pmatrix} \xrightarrow[n\to\infty]{\text{d}} \mathcal{MN}_{s \times B}\left(0, I, \Xi\right)$$

where $\mathcal{MN}$ denotes the matrix normal distribution. Hence the B vectors of effective scores $(S_1^1, \dots, S_K^1), \dots, (S_1^B, \dots, S_K^B)$ converge to i.i.d. random vectors.

For each k, the standardized scores $\tilde{S}_k^b$ are obtained dividing the effective scores S_k^b by their standard deviation $\operatorname{var}(S_k^b \,|\, F^b)^{1/2}$, as in (4). De Santis et al. (2022) show that these standard deviations are asymptotically independent of b (see the proof of Theorem 2). Therefore, the B vectors of the absolute values of standardized scores $(|\tilde{S}_1^1|, \dots, |\tilde{S}_K^1|), \dots, (|\tilde{S}_1^B|, \dots, |\tilde{S}_K^B|)$ converge to i.i.d. random vectors. As a consequence, the combinations of their elements $T^1, \dots, T^B$ defined

in (6) converge to i.i.d. random variables. Moreover, for each variable k, high values of $|\tilde{S}_k^1|$ correspond to evidence against $\mathcal{H}_k$ and ψ is non-decreasing in each argument, and so high values of T^1 correspond to evidence against $\mathcal{H}$. From Hemerik et al. (2020) (see Lemma 1),

$$\lim_{n\to\infty} P\left(T^1 > T^{(\lceil(1-\alpha)B\rceil)}\right) = \frac{\lfloor \alpha B \rfloor}{B} \leq \alpha.$$

Finally, consider the more general case where we also allow for specifications that change the sample size, so that the response vector Y_k may vary between models and have different lengths. The proof is written analogously to the previous one, with a slight modification of the sign-flipping matrices within each model. In model k we use F_k^b, which is obtained from F^b by removing the diagonal elements corresponding to the removed observations.

Theorem 2 gives an asymptotically exact test for the global null hypothesis $\mathcal{H}$ that the coefficient of interest is null in all considered models. A global p-value can be obtained directly as

$$p = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{T^b \geq T^1\}$$

(Hemerik and Goeman, 2018).

An important role is played by the choice of the function ψ that combines the flipped standardized scores to define the global test statistic (6). There is a plethora of possible choices, each of them having different power properties in different settings. The most intuitive choices are the mean

$$T^b_{\text{mean}} = \frac{1}{K}\sum_{i=1}^{K} |\tilde{S}_k^b| \qquad (b = 1, \ldots, B) \tag{7}$$

and the maximum

$$T^b_{\text{max}} = \max_k |\tilde{S}_k^b| \qquad (b = 1, \ldots, B) \tag{8}$$

but the definition of the test remains flexible and general, allowing for several combinations. Other possible global test statistics can be obtained transforming the standardized scores $\tilde{S}_k^b$ in p-values p_k^b, and then considering p-value combinations. The p-values can be defined either

through parametric inversion of the scores or using ranks; we suggest this second choice, where

$$p_k^b = \frac{1}{B}\sum_{c=1}^{B} \mathbf{1}\{|\tilde{S}_k^c| \geq |\tilde{S}_k^b|\} \qquad (k = 1, \ldots, K;\ b = 1, \ldots, B).$$

Subsequently, the p-values can be combined with different methods such as those described and compared in Pesarin (2001). We mention especially Fisher (1925)

$$T^b_{\text{Fisher}} = -2\sum_{k=1}^{K} \log p_k^b \qquad (b = 1, \ldots, B) \tag{9}$$

and Liptak/Stouffer (Liptak, 1958)

$$T^b_{\text{Liptak}} = -\sum_{k=1}^{K} \zeta(p_k^b) \qquad (b = 1, \ldots, B)$$

where $\zeta(\cdot)$ denotes the quantile function of the standard normal distribution.

3.2. Post-selection Inference

In the previous section, we considered different plausible specifications of a GLM and defined the global null hypothesis $\mathcal{H}$ that a predictor of interest does not influence the response in any of these models. We constructed a test that combines the models' standardized scores to test $\mathcal{H}$ at the level α, and thus ensuring weak control of the FWER. Therefore, if $\mathcal{H}$ is rejected, we can state with confidence $1-\alpha$ that there is at least one model in which the predictor of interest has an influence on the response variable. In this section, we show that the global test statistic T^b defined in (6) can be used to make additional inferences about the models in two ways. We rely on the closed testing framework (Marcus et al., 1976), which has been proven to be the optimal way to construct multiple testing procedures, as all FWER, TDP, and related methods are either equivalent to it or can be improved by it (Goeman et al., 2021). It is based on the principle of testing different subsets of hypotheses by means of a valid local test at the α level, which in this case is the test of Theorem 2.

First, to obtain adjusted p-values for each individual model, we apply the maxT method of Westfall and Young (1993), which corresponds to using as a global test statistic the maximum defined in (8). This procedure provides for a dramatic shortcut of the closed testing framework is

fast and feasible even for high values of K and B. The resulting p-values are adjusted for multiplicity, ensuring strong control of the FWER. Researchers can postpone the choice of the preferred model after seeing the data, while still obtaining valid p-values. Used in this way, the method allows researchers to make selective inferences. Where selective inference is a possible cause of the replication crisis when error rates are not controlled (Benjamini, 2020), the PIMA procedure provides strong FWER control, allowing researchers to select a model after analyzing a multiverse of models without inflating the risk of a false positive.

Second, we can construct a lower $(1-\alpha)$-confidence bound for the proportion of models where the coefficient is non-null (TDP), using the general framework of Genovese and Wasserman (2006) and Goeman and Solari (2011) or, when the combining function ψ can be written as a sum, the shortcut of Vesely et al. (2023). The method allows one to compute a confidence bound for the TDP not only for the whole set of models, but also simultaneously over all possible subsets without any adjustment of the α level. Simultaneity ensures that the procedure is not compromised by selective model selection. In this framework we are not able to individually identify statistically significant models, but in some cases, reporting the TDP may be more powerful than individually adjusted p-values.

To conclude, the PIMA approach allows researchers to make selective inference on the parameter of interest in the multiverse of models, providing not only a global p-value but also individually adjusted p-values and lower confidence bounds for the TDP of subsets of models. The PIMA procedure is exact only asymptotically in the sample size n; despite this, we will show through simulations that it maintains good control of the Type I error even for small values of n. Furthermore, as shown in the real data analysis of Section 5, the same inference framework can be trivially extended to the case where we are interested in testing multiple parameters, i.e. where β is a vector. Analogous to the extension from a single model to the multiverse, it is sufficient to define global test statistics (6) for all individual parameters of interest using the same random sign-flipping transformations.

3.3. Comparing PIMA with other proposals

In this section we discuss and evaluate possible competitors to the PIMA procedure to test the global null hypothesis $\mathcal{H}$. A first naive approach would be to rely on a parametric method. However, after computing a test for each model, there is the need to combine the univariate tests into a multivariate one. Since these tests coming from different specifications are generally not independent and their dependence is very difficult to model formally, the safest option is to use a Bonferroni correction. This approach has the invaluable advantage of simplicity, but has very low power in practice. This is mainly due to the strong correlation between model estimates that usually occurs when different specifications of the same model are tested.

As mentioned in Section 1, the specification curve analysis of Simonsohn et al. (2020) represents a first attempt to cast the descriptive approach of multiverse analysis into an inferential framework. Two approaches are proposed. The first one relies on a naive permutation of the tested predictor followed by a re-fitting of the models; the subsequent combination of the test statistics of each model follows the same logic exposed in Section 3.1. This method is only valid when the predictors are orthogonal, a setting that is typically limited to fully balanced experimental designs. Hence the method is no longer valid neither in experimental designs with unbalanced levels nor in non-experimental designs. The second approach can be used in the more general case of non-experimental settings. It was originally defined for LMs and is based on the bootstrap method of Flachaire (1999). For each specification, the model with the observed data is fitted (i.e., $y_i = \beta x_i + \gamma z_i + \varepsilon_i$), producing the estimates of the parameters β and γ. Then, a null response $\dot{y}_i$ is generated by subtracting the estimated effect of the predictor of interest x_i on y_i: $\dot{y}_i = y_i - \hat{\beta} x_i = (\beta - \hat{\beta}) x_i + \gamma z_i + \varepsilon_i$, where $\hat{\beta}$ is the sample estimate of β. The random variable $\beta - \hat{\beta}$ has zero-mean, therefore a null distribution of $\hat{\beta}$ can be obtained by re-fitting the model on bootstrapped data $\dot{y}_i,\ x_i,\ z_i$. The resulting bootstrapped distribution of $\hat{\beta}$ is used to compute the p-value for $\mathcal{H}$. Subsequently, the same resampling scheme is applied to each specification, and the resulting p-values are merged through appropriate combinations: the median, Liptak/Stouffer (Liptak, 1958), and the count of specifications that obtain a statistically significant effect. In the case of LMs, both the bootstrap method and PIMA are robust to heteroscedasticity (Flachaire,

1999) and ensure asymptotic control of the Type I error. However, while the univariate test of the bootstrap method is still only asymptotically exact, the sign-flip score test on which PIMA relies has exact univariate control. Finally, the bootstrap refits the model at each step, and so requires a substantially larger computational effort, as will be confirmed in simulations.

The same bootstrap procedure of Simonsohn et al. (2020) is then extended to the case of GLMs. However, this extension is not always valid in our view. For GLMs, the authors base the bootstrap on the definition of null responses of the form $\dot{y}_i = g^{-1}(g(y_i) - \hat{\beta}x_i)$, but this proposal turns out to be very problematic for some models. For instance, the same issue occurs when considering the binomial logit-link model, where $y_i \in \{0, 1\}$ and $\dot{y}_i = \text{expit}(\text{logit}(y_i) - \hat{\beta}x_i)$. When the response is $y_i = 0$, we have that $\text{logit}(0) = -\infty$ and so the null response is always $\dot{y}_i = 0$, regardless of the value of $\hat{\beta}x_i$. Similarly, when $y_i = 1$ we always obtain $\dot{y}_i = 1$. This means that the effect of the tested variable is never removed when computing the null response, contrary to what happens in the case of the LM for which the method was originally defined. The same problem arises for other GLMs that lead to infinite values, such as the Poisson log-link model, for which $\dot{y}_i = \exp(\log(y_i) - \hat{\beta}x_i)$ is always 0 when $y_i = 0$. Thus, in the case of GLMs we discourage the use of the proposal of Simonsohn et al. (2020). The poor control of the Type I error in practice will be shown in the simulation study of Section 4.

Finally, specification curve analysis only explores weak control of the FWER, i.e., infers on the presence of at least one significant specification. We underline the importance of the post-selection inference step introduced in PIMA, which makes it possible to determine which models have a significant effect, and thereby allows researchers to gain a better understanding of the overall analysis.

4. Simulations

The following simulation study aims to assess the control of Type I error and to quantify the power of the global test of Theorem 2, by performing a comparison with the bootstrapped method of specification curve analysis (Simonsohn et al., 2020). Adjusted p-values for individual specifications are not reported, since Type I error control of the global test automatically ensures FWER control through the closed testing principle, as argued in Section 3.2.

We set a common framework for all simulations, based on the settings used for specification curve analysis. Simonsohn et al. (2020) simulated data generating the response Y from a latent variable X^ℓ through a GLM, then considered a multiverse analysis with five different models. Each model uses a different predictor X_k, which is taken as a proxy for the latent X^ℓ and is generated to be strongly correlated with it. We extend this setting by adding a confounder Z that is also correlated with X^ℓ. The pipeline for the analysis is as following. First, we simulate the latent variable X^ℓ and the confounder Z from a multivariate normal distribution with mean zero, unitary variance and covariance $\rho_{X^\ell Z} = 0.6$. Then we generate the response Y through a GLM, taking

$$g(\mu_i) = x_i^\ell \beta + z_i \gamma + \gamma_0.$$

Finally, we consider a multiverse analysis with five models. For each model k, we generate a new predictor X_k so that it has a correlation with the latent variable $\rho_{X^\ell X_k} = 0.85$. Then we fit a GLM with X_k as the predictor of interest and Z as the confounder.

We consider four scenarios, where in the first three we fit the correct model, while in the last scenario the variance in the fitted model is misspecified as follows:

1. LM with homoschedastic Gaussian errors: $\gamma_0 = 0$, $\gamma = 2$, $\beta = 0$ (under the null hypothesis) or $\beta = 0.2$ (under the alternative hypothesis), homoscedastic normal errors with variance 1;
2. Binomial logit-link model: $\gamma_0 = 0$, $\gamma = 2$, $\beta = 0$ (under the null hypothesis) or $\beta = 0.5$ (under the alternative hypothesis);
3. Poisson log-link model: $\gamma_0 = 0$, $\gamma = 2$, $\beta = 0$ (under the null hypothesis) or $\beta = 0.08$ (under the alternative hypothesis);
4. Data are generated with a Negative Binomial log-link model: $\gamma_0 = -2$, $\gamma = 2$, $\beta = 0$ (under the null hypothesis) or $\beta = 0.25$ (under the alternative hypothesis), and dispersion parameter $\theta = \mu$ (so that the variance $\mu + \mu^2/\theta = 2\mu$ is twice the variance expected in a Poisson model). In this case a Poisson log-link model is fitted. In this scenario, we evaluate the robustness of the methods to misspecification of the variance.

For each scenario, we apply different tests with the scope to assess both the Type I error rate

and the power, setting the coefficient of interest β to 0 in the first case (null hypothesis $\mathcal{H}$) and at non-null values in the second (alternative hypothesis). We start exploring the behavior of univariate tests, applying, for each of the five models, three different methods: the sign-flip score test of Theorem 1, the bootstrapped method of the specification curve analysis (Simonsohn et al., 2020), and a suitable parametric test (t-test for LM, Wald test for the other GLMs). Subsequently, we then combine the information derived from the five considered models. We apply the PIMA method, taking as global test statistic the mean (7) and the maximum (8). We also report results for the bootstrapped method (Simonsohn et al., 2020), combining the individual specifications' p-values with Stouffer and the median. We do not consider the combination that counts the specifications having a statistically significant effect since it implicitly involves one-sided alternatives, and dealing with the control of directional errors in multiple testing is a non trivial task (Shaffer, 1980; Finner, 1999) which deserves the effort of a more formal dissertation. Finally, an additional parametric global test could be obtained from the Bonferroni combination of the five univariate parametric tests; however, this is not feasible in practice, as the Bonferroni method results to be extremely conservative.

Throughout the simulations, we vary the sample size $n \in \{100, 250, 500\}$. Furthermore, we use $B = 250$ random sign-flipping transformations and bootstraps; the choice of using a relatively small number is motivated by the huge computational cost required by the bootstrap, which refits the model at each step. We remark that the number of the random resamplings – bootstraps or sign-flips – does not affect the control of the Type I error (Hemerik and Goeman, 2018; Ramdas et al., 2023). Each scenario is simulated 5,000 times. This implies a standard error around significance level 5% equal to $\sigma_{\text{err}} = \sqrt{0.05 \cdot 0.95/5000} = 0.003$, and so the bounds in this case are $\alpha \pm 1.96\sigma_{\text{err}} = (0.044, 0.056)$.

Figure 3 reports the Type I empirical error rates for the different methods in the four scenarios. Each row reports the rejection proportion under the null hypothesis $\mathcal{H}$ ($\beta = 0$) for the five univariate models. Under the linear model (top-left plot), parametric tests and flipscores ones show a perfect control of the Type I error as expected from the theory. Even the bootstrap method shows optimal behavior, although the control is ensured only asymptotically (Freedman, 1981). In the Binomial (top-right) and Poisson (bottom-left) scenarios, the parametric and the

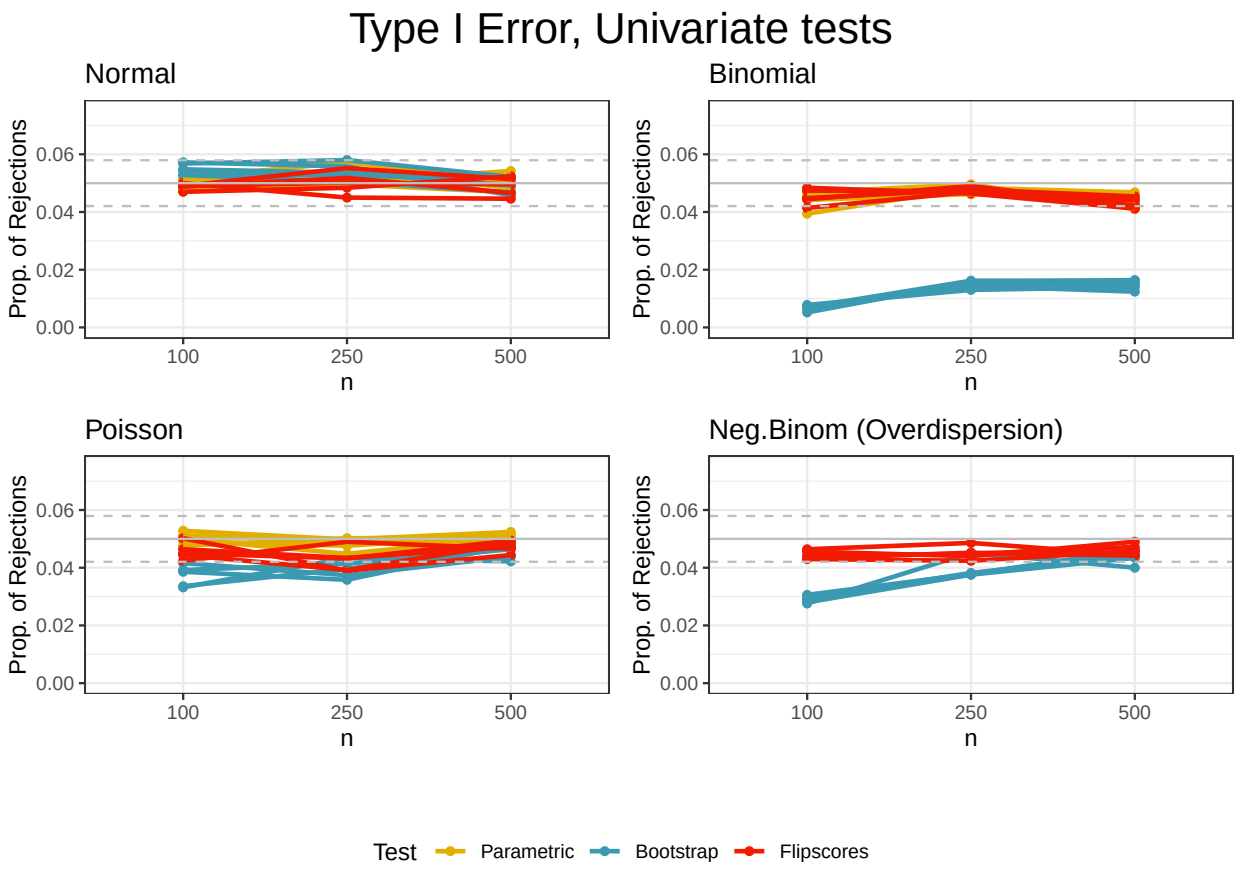

FIGURE 3.

Simulations (univariate) under the null hypothesis $\mathcal{H} : \beta = 0$: empirical Type I error of different methods for sample size $n \in \{100, 250, 500\}$ under four scenarios. For the Neg. Binom scenario the empirical Type I errors of the bootstrap approach exceed the upper limits of the ordinates – ranging between 0.154 and 0.170 – and are not shown. The dotted horizontal lines around 0.05 correspond to the 95% simulation error's limits.

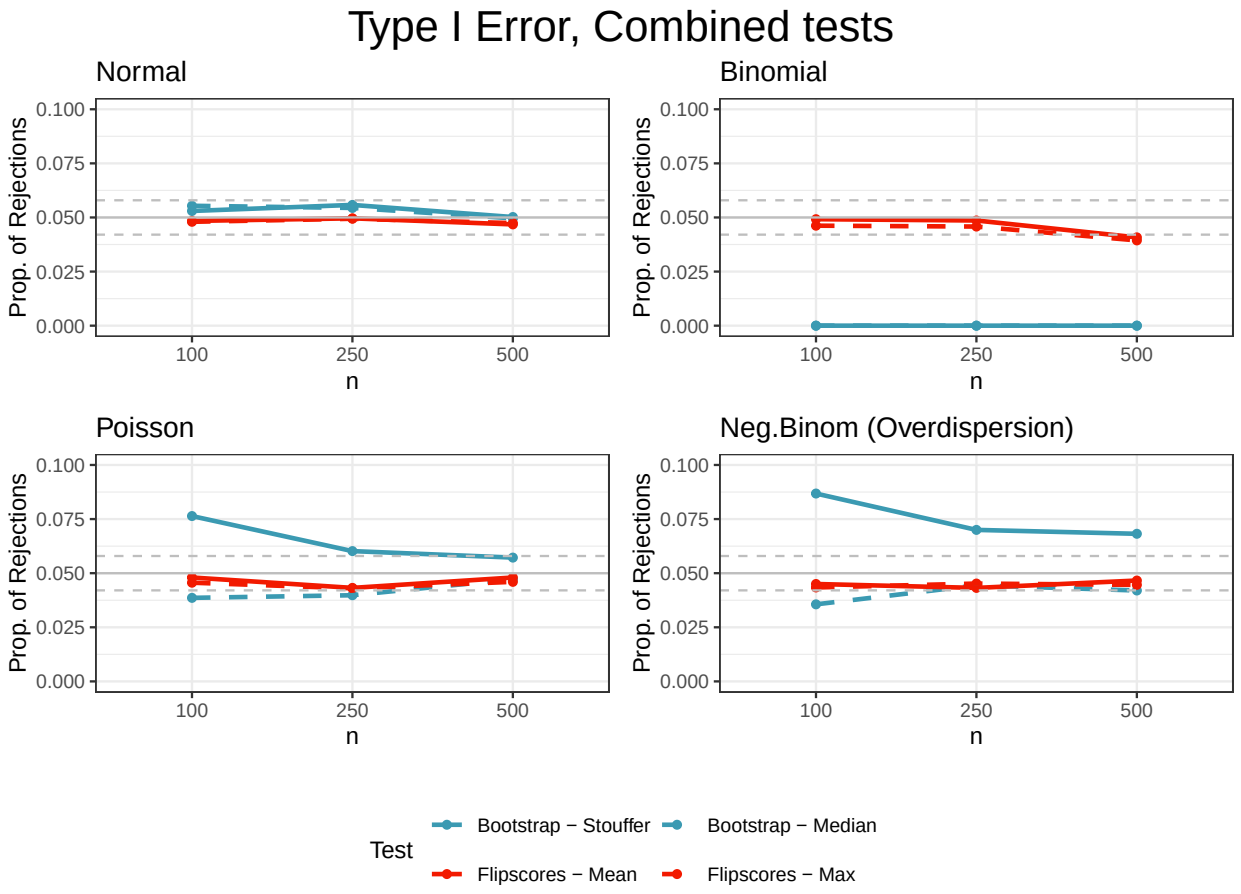

FIGURE 4.

Simulations (combined) under the null hypothesis $\mathcal{H} : \beta = 0$: empirical Type I error of different methods for sample size $n \in \{100, 250, 500\}$ under four scenarios. The dotted horizontal lines around 0.05 correspond to the 95% simulation error's limits.

flipscores are formally proven to have an asymptotic control of the Type I error; the simulation confirms the good control in practice. The bootstrap approach shows a conservative behavior in these settings, especially for small sample sizes. Finally, in the Negative Binomial scenario, where a Poisson model with log link is fitted (bottom-right), the parametric model largely exceeds the putative $\alpha = 0.05$ level, ranging between 0.154 and 0.170; this is not reported in the figure merely for graphical reasons. In the same setting, the flipscores test performs well while the bootstrap remains conservative.

Figure 4 shows the results of the multiverse of models under $\mathcal{H}$. The boostrap and flipscores methods offer a comparable level of control for the linear model scenario. For GLMs, the boostrap does not seem to adequately control the Type I error, resulting too conservative in the Binomial scenario and exceeding the nominal level of 5% in most cases for the Poisson and Negative Binomial scenarios.

Considering only the methods controlling for the Type I error in the previous univariate tests, the power increases as the sample size increases (Figure 5); a slightly higher performance was observed for the parametric and flipscore procedures compared to the boostrap test in the Binomial scenario, and for the parametric method in the Poisson case. In the combined tests, the performance of the boostrap method (only with a Stouffer aggregation) is higher than the flipscores, while in the Binomial scenario the boostrap method fails. For the Poisson and the Negative Binomial scenario, the simulations offer a similar power for both methods.

In conclusion, these simulations provide some insights to evaluate as the PIMA approach provides a general framework for the multiverse analysis, while the validity of the bootstrap method offers a limited superiority only under the scenario with linear model and Gaussian error. An exhaustive analysis in a wider range of settings would be of great interest, but would be very extensive. Indeed, the PIMA method is extremely general and flexible, since it can be applied to any GLMs. A substantial number of scenarios could potentially be explored, considering different combinations of the characteristics studied here, as well as many others, such as the total number of predictors, their covariance, the nuisance parameters γ, the number and Type of specifications, etc. Consequently, such an analysis is left for future work.

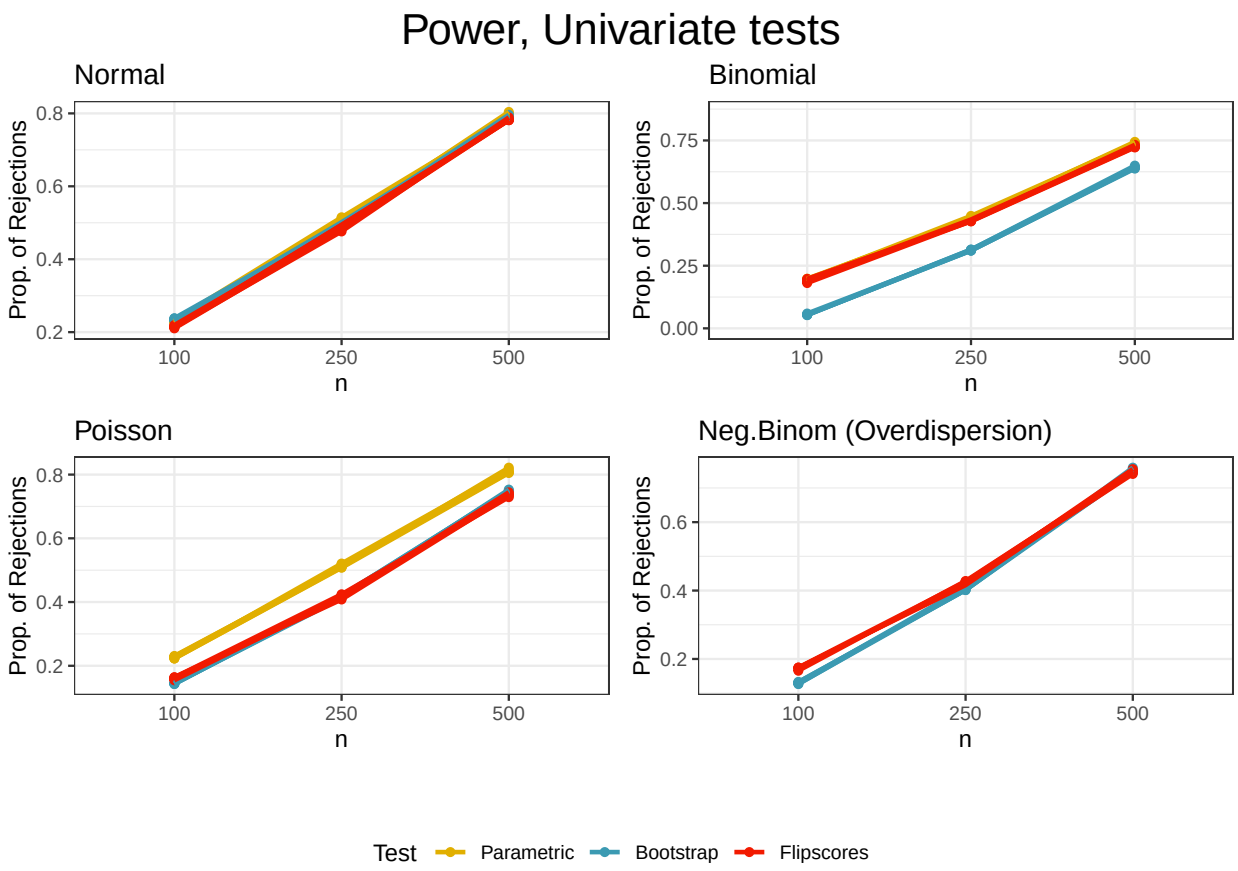

FIGURE 5.

Simulations (univariate) under the alternative hypothesis ($\beta = 1$): empirical power of the methods controlling for Type I error for sample size $n \in \{100, 250, 500\}$ under four scenarios. The power of the methods that do not control the Type I error – i.e. exceeding the upper limit in the respective setting in Figure 3 – are not shown.

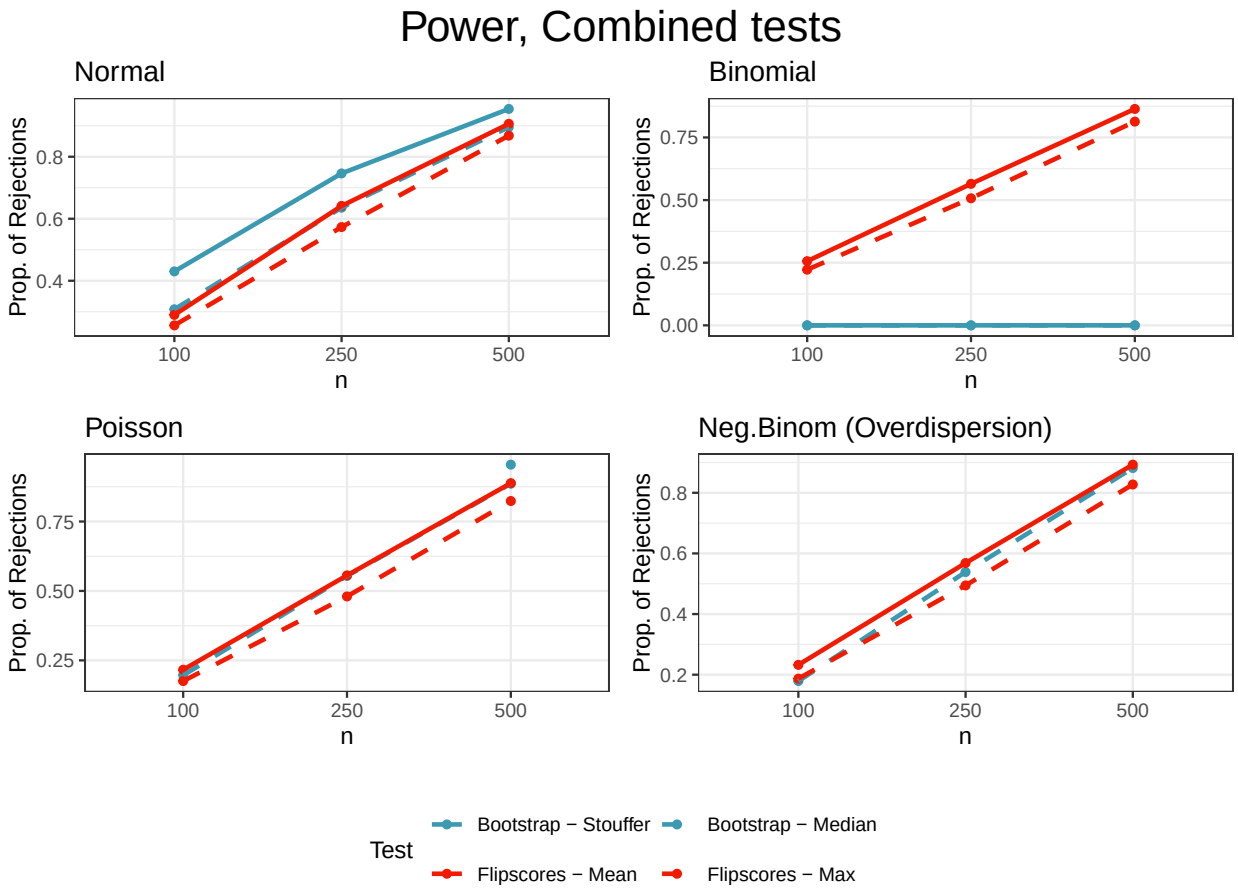

FIGURE 6.

Simulations (combined) under the alternative hypothesis ($\beta = 1$): empirical power of the methods controlling for Type I error for sample size $n \in \{100, 250, 500\}$ under four scenarios. The power of the methods that do not control the Type I error – i.e. exceeding the upper limit in the respective setting in Figure 4 – are not shown.

5. Data analysis: The COVID-19 vaccine hesitancy dataset

5.1. Description of the dataset

The COVID-19 vaccine hesitancy dataset collected information on people's intention to get vaccinated, sociodemographic characteristics and other important variables, i.e. the perceived risk related to the COVID-19 contagion, doubts about vaccines, and conspiracy (Caserotti et al., 2021). This survey was the first data collection that included data on vaccine hesitancy before, during and after the lockdown in Italy, which lasted from March 8 until May 3, 2020. The dataset is formed by a collection of voluntary respondents on the basis of a snowball sampling technique. The willingness to be vaccinated was originally collected on a scale between 1 and 100; in this example, we mark as hesitant all people with an index below 100 ($n = 1359$), while the others are marked as not hesitant ($n = 909$). The main characteristics are reported in Table 1, in general and by the state of reluctance. Three variables were marginally associated with the status of hesitancy: calendar period, perceived risk of COVID-19, and doubts about vaccines.

5.2. Inferential approach

We want to assess whether the doubts of the people about a potential vaccine against COVID-19 remained constant or reported a substantial change before, during and after the Italian lockdown, due to different perceptions of risk associated with COVID-19 contagion during the different phases of the epidemic outbreak. To estimate the adjusted effect of the calendar period, several confounders are taken into account: `Covid_perc_risk`, COVID-19 Perceived risk, a scale defined combining different COVID-19 risk subscales (for further details, see Caserotti et al. (2021)); `doubts_vaccine`, vaccine doubts on a 0-100 scale; `Age`, age in years; `Gender`, gender; `Age*Gender`, interaction between age and gender; `deprivation_ index`, Italian Deprivation Index at the city of residence; `geo_are`, geographical area.

The variable to be tested is the date of the period of data collection `Period` recoded in a categorical variable with three levels according to the temporal window of the Italian lockdown: pre-lockdown (`Pre`), during (`Lockdown`) and post-lockdown (`Post`). We are interested in all three possible comparisons, and their post-hoc corrected p-values. For each comparison, we fit a model

Characteristic	**Overall**	**Hesitant**	**No hesitant**	***P*-value**[2]
	N = 2,268[1]	N = 1,359[1]	N = 909[1]	
Gender				0.065
Female	1,585 (70%)	930 (68%)	655 (72%)	
Male	683 (30%)	429 (32%)	254 (28%)	
Age (years)	35 (26, 49)	35 (27, 48)	35 (25, 51)	0.4
Geographical area				0.2
Center	95 (4.2%)	65 (4.8%)	30 (3.3%)	
North	2,015 (89%)	1,200 (88%)	815 (90%)	
South	158 (7.0%)	94 (6.9%)	64 (7.0%)	
Period				< 0.001
Pre-lockdown	845 (37%)	609 (45%)	236 (26%)	
Lockdown	978 (43%)	494 (36%)	484 (53%)	
Post-lockdown	445 (20%)	256 (19%)	189 (21%)	
COVID-19 Perceived Risk	123 (80, 162)	103 (62, 146)	149 (110, 176)	< 0.001
Vaccine doubts	8 (0, 25)	11 (3, 40)	2 (0, 10)	< 0.001
Deprivation Index	-0.69 (-1.61, 0.43)	-0.69 (-1.64, 0.43)	-0.69 (-1.49, 0.43)	0.6

[1] *n (%); Median (IQR)*

[2] *Pearson's Chi-squared test; Wilcoxon rank sum test.*

TABLE 1.

COVID-19 vaccine hesitancy: variables included in the analysis, overall and by hesitancy status.

with a zero-centered contrast that models the comparison of interest. For example, to test the difference between `Post` and `Pre` we define X as a variable with a value of 1 for `Post`, -1 for `Pre` and 0 for `Lockdown`. The confounders Z comprise a dummy variable for the level not involved in the comparison together with the above-mentioned confounders.

Having a dichotomous response $Y = \{\text{not hesitant, hesitant}\}$, then recoded as $Y = \{1, 0\}$, we use a GLM model with binomial response and logit link:

$$y_i \sim Bernoulli(p_i), \qquad p_i \in (0, 1)$$
$$g(p_i) = \log \frac{p_i}{1 - p_i} = \alpha + \beta x_i + \gamma z_i.$$

In order to implement a flexible approach, the relationship of the continuous predictors with the response is modeled also by basis of splines (B-splines). For each continuous predictors `Covid_perc_risk`, `doubts_vaccine`, `deprivation_index` and `Age`, three transformations are tested: the natural variable, as well as a B-spline with three and four degrees of freedom. In total, there are $K = 3^4 = 81$ model specifications. For each comparison, e.g. `Post - Pre`, the k-th tested null hypothesis in model k is defined as $\mathcal{H}_k^{Post-Pre}: \ \beta_k^{Post-Pre} = 0$. The global null hypothesis is the intersection of all null hypotheses, $\mathcal{H}^{Post-Pre} = \cap_k^K \mathcal{H}_k^{Post-Pre}$.

For each comparison we apply the PIMA framework with the max-T combining function. Thus, we obtain: 1) a global p-value for the null hypothesis of no change over time (weak control of the FWER); 2) adjusted p-values for all individual models (strong control of the FWER); 3) lower confidence bounds for the TDP, i.e, the minimum proportion of models that show a significant difference. Furthermore, in this peculiar case we need to jointly test all possible pairwise comparisons: $\mathcal{H}^{Post-Pre} \cap \mathcal{H}^{Post-Lockdown} \cap \mathcal{H}^{Lockdown-Pre}$. Accounting for the 81 model specifications, each with three possible comparisons, we obtain 243 tests in total. The solution to this inferential problem is natural in the PIMA framework, as it is sufficient to define the closure set as the closure of the union of the univariate hypotheses of the three comparisons.

5.2.1. Results

We first report results for a parametric binomial model with linear predictors (i.e., natural variables, no B-spline used here) and two 0-centered contrasts variables that model the three-level

`Period` variable. Table 2 reports the summary, while Table 3 shows the post-hoc Tukey correction for the three pairwise comparisons.

	Estimate	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	-2.069	0.264	-7.843	0.000
Pre-lockdown	-0.291	0.079	-3.673	0.000
Lockdown	-0.089	0.072	-1.236	0.216
Vaccine doubts	-0.036	0.003	-13.029	0.000
Deprivation Index	-0.011	0.028	-0.389	0.697
COVID-19 Perceived risk	0.015	0.001	12.914	0.000
Age (+1 years)	0.012	0.004	2.665	0.008
Gender [males]	0.539	0.297	1.811	0.070
Geo. Area [North]	-0.026	0.176	-0.149	0.881
Age*Gender [males]	-0.016	0.007	-2.106	0.035

TABLE 2.

Summary of the estimated logistic regression model with logit link and linear confounders for COVID-19 vaccine hesitancy dataset. Period reference category is Post-lockdown.

	Coefficients	**Sigma**	**Tstat**	***p*-values**
Lockdown - Pre-lockdown	0.202	0.124	1.634	0.230
Post-lockdown - Lockdown	0.469	0.138	3.392	0.002
Post-lockdown - Pre-lockdown	0.671	0.150	4.476	0.000

TABLE 3.

Post-hoc pairwise comparisons with (Tukey) correction of the logistic regression model with logit link and linear confounders.

As introduced in the previous section, the multiverse analysis framework is built on the basis of three possible transformations for each continuous predictor (81 models) and also across the 3 comparisons (`Pre - Lockdown`, `Pre - Post`, `Lockdown - Post`), leading to a multiverse of $81 \cdot 3 = 243$ models. Figure 7 reports the results visually, while detailed results are reported in

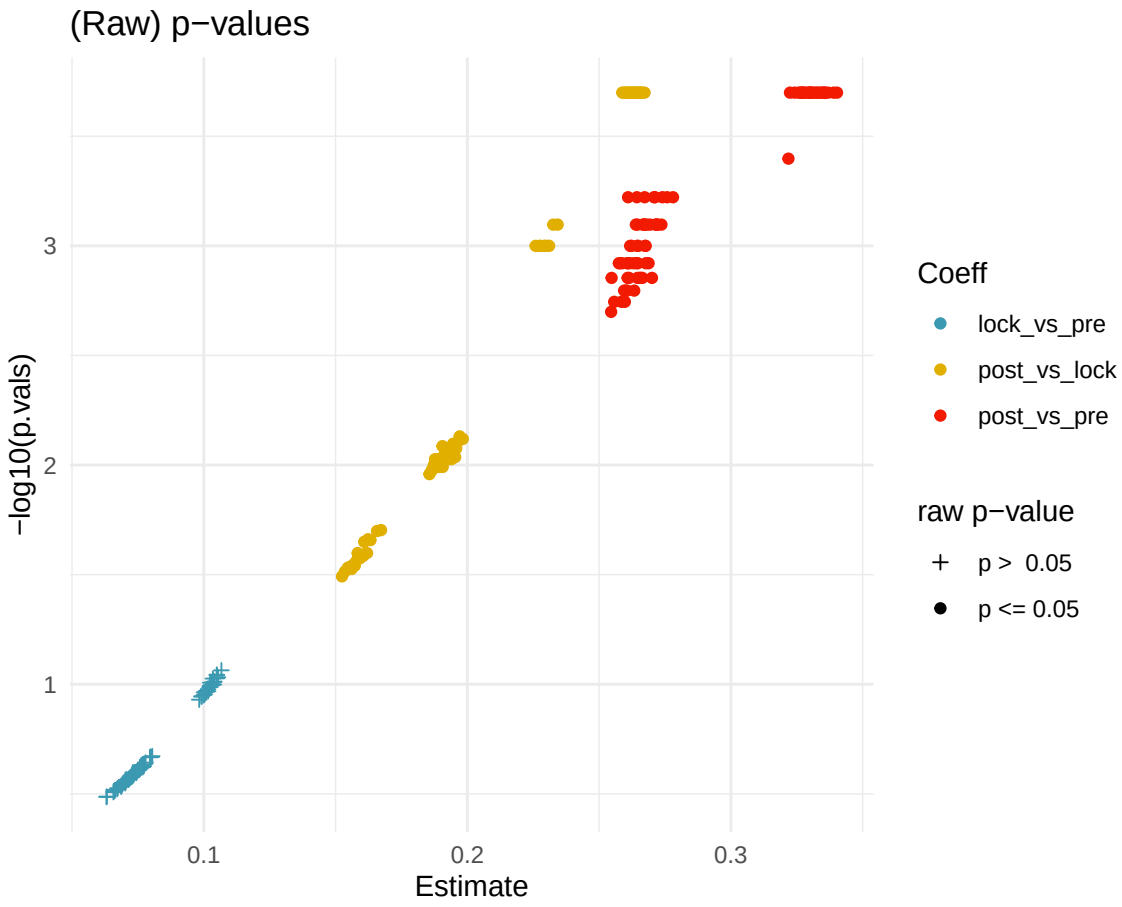

FIGURE 7.

Raw p-values vs. coefficient estimates of the post-hoc comparisons of 81 tested logistic models in the PIMA of COVID-19 vaccine hesitancy dataset.

the Appendix. The usual descriptive interpretation of a multiverse analysis allows us to observe descriptively that the yellow and red clusters of tests yield p-values smaller than 0.05, but we cannot claim that these results are statistically significant due to the possibility that such a claim would have an unacceptably high false positive rate.

We now move from the descriptive to the inferential analysis. The global test with post-hoc correction is shown in Table 4. The comparisons `Post - Pre` and `Post - Lockdown` are significant (overall, over the 81 models of the multiverse), while the comparison `Lockdown - Pre` is not. We point out that the post-hoc correction is based on the three comparisons, and each comparisons is

based on the combination of the 81 models. For example, claiming that the comparison `Post - Pre` is significant allows us to account only for the fact that at least one of the 81 models has non-null coefficient for this comparison (and assuming that all models are correctly specified). In order to select the most promising models among these, we need to shift the multiplicity correction to the levels of each individual model. This is done in Figure 8, where the adjusted p-values are reported (the table with the detailed results is reported in the supplementary material).

Coeff	**Stat**	**nMods**	**S**	**Pr(>|z|)**	**p.adj (post-hoc)**
Lockdown - Pre-lockdown	maxT	81	1.71	0.1396	0.1396
Post-lockdown - Lockdown	maxT	81	3.78	0.0008	0.0008
Post-lockdown - Pre-lockdown	maxT	81	4.43	0.0002	0.0006

TABLE 4.

Pairwise comparisons of the maxT global test between periods with post-hoc correction in the PIMA of COVID-19 vaccine hesitancy dataset.

Finally, Table 5 reports the number of true discoveries and the TDP for each comparison. For the `Post - Pre` comparison, all models show a significant difference (after multiplicity correction), while those in the `Lockdown - Pre` comparison showno significant effect. The comparison `Post - Lockdown` has an intermediate number of significant comparisons ($29/81 = 36\%$).

Coeff	**Stat**	**nMods**	**True Discoveries**	**Proportion**
LockDown - Pre	maxT	81	0	0%
Post - LockDown	maxT	81	29	36%
Post - Pre	maxT	81	81	100%

TABLE 5.

Lower 0.95-confidence bound for the number of true discoveries in each comparison in the PIMA of COVID-19 vaccine hesitancy dataset.

6. Conclusion and final remarks

In this paper we propose PIMA, a formal inferential framework to multiverse analysis (Steegen et al., 2016). Our approach allows researchers to move beyond a descriptive

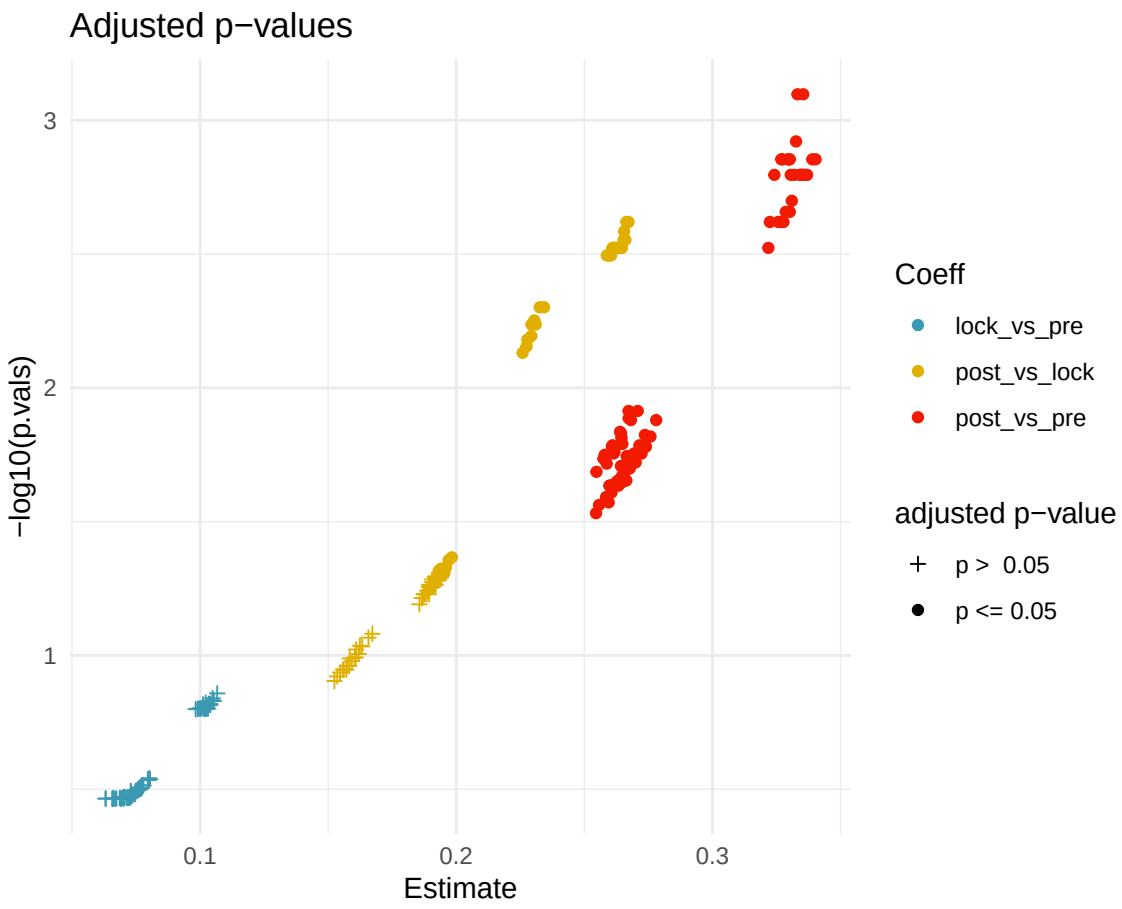

FIGURE 8.

Adjusted p-values vs. coefficient estimates of the post-hoc comparisons of 81 tested models in the PIMA of COVID-19 vaccine hesitancy dataset.

interpretation of the results of a multiverse analysis, and extends other methods to summarize the multitude of performed analyses, such as specification curve (Simonsohn et al., 2020) and vibration analysis Klau et al. (2020) to any generalized linear model. By extending the sign-flip score test (Hemerik et al., 2020; De Santis et al., 2022) to the multivariate framework, researchers can now make use of the full variety of multivariate and multiple testing methods based on conditional resampling to obtain: 1) weak control of the FWER to test if there is an "overall" effect in one of the models explored in the multiverse analysis; 2) strong control of the FWER (i.e., adjusted p-values for each tested model) that allows the researcher to select the models that show a significant effect; 3) a lower confidence bound for the proportion of true discoveries among the tested models. Furthermore, PIMA proves to be very robust to over/under-dispersion, allowing for a wide range of models and possible data preprocessing.

This flexibility, however, does not exempt the researcher of the responsibility of the analysis. Some further remarks and considerations should be made in this regards.

Define your theoretical model. In the multiverse analysis framework, each specification should follow a model that is based on a strong theory developed within a research field (e.g., psychology, medicine, physics, etc.). As an example, in epidemiology, a researcher usually defines a set of variables called "confounders" in order to adjust for the estimated effect between the dependent variable (outcome) and the independent variable (determinant) in a quasi-experimental design. In this case, any specification can be plausible if it includes the same set of confounders as an evaluation of the same initial theoretical model. Exclusion of some confounders is commonly used in sensitivity analysis, but it could lead to implausible specifications because of the potential mismatch with the theoretical initial model.

Plan your analysis in advance. It is important to note that the PIMA method is not iterative, i.e. the analysis specifications must be planned before performing the multiverse analysis. Failure to do so (i.e. adding or removing models after seeing the results) will add a layer of data manipulation, which is impossible to model and hard to formalize, and therefore can inflate the Type I error rate. The multiverse approach allows the researcher to plan (in advance) a plethora of models to explore, rather than just a single pre-specified model. However, it is recommended to pre-register PIMA before it is performed.

Be parsimonious. There is virtually no limit to the number of models that can be used, as the proposed PIMA approach will integrate all the resulting information. However, the power will be affected by these choices. Indeed, the overall power to find a significant effect depends on the power of each individual model. Although adding "futile" models will not decrease the quality of false positive control, it will decrease the power of the global test and therefore the ability to detect significant effects.

Be exhaustive. There is a further consideration that applies not only to our inferential methods but also to descriptive methods such as multiverse analysis, specification curve and vibration analysis (and to data modeling, more broadly). When planning the data transformations, the practitioner must realize that failing to take into account any relationship between confounders and the response variable may be a catastrophic source of false positive results. This case is very well covered in any basic course in statistical modeling, but it may be useful to provide a flavor of the consequences of an inaccurate choice of models in the analysis in practice. We run a simple simulation under the same linear homoscedastic normal framework described in Section 4. In this case, however, we do not include the last two confounders in any of the models. The empirical Type I error rate exceeds 0.30 (nominal level $\alpha = 0.05$) in all tested models. As a consequence, the combined model exceeds the nominal level by the same amount. The same behavior can be seen in the parametric approach. As practical advice, we recommend including all potential confounders in all models, since losing control of the Type I error in any of them will make the inference unreliable.

A more subtle but very relevant example is the case where some transformation of the confounders does not account for all the dependence between them. For instance, suppose that a confounder Z has a linear relationship with the response Y and with the variable of interest X. Now, to account for non-linear effects, the researcher decides to use a median-split transformation of Z. The resulting test on the coefficient of X will lose its control of the Type I error. To elucidate this case in practice, we run a second simulation, again under the same setting described above (linear homoscedastic model). In this case, we include all the confounders, but we use a median-split transformation instead of the parabolic models. With sample size $n = 200$, the empirical Type I error of the true (linear) model is under control (sign-flip score test: 0.051,

parametric: 0.054), while it largely exceeds 5% for any other model that median-splits the predictors, reaching 0.211 for the sign-flip score test (and 0.219 for the parametric test) when the model has all the three confounders with a median-split. As a direct consequence of the loss of control of Type I error of the univariate models, the PIMA method loses its error control as well. The empirical Type I error is 0.180 for the maximum, and similar for other combining functions. It would be easy to define more dramatic scenarios, of course.

Thorough discussion of the results. The previous consideration may be uncomfortable. It implies that every significant test must be evaluated with great care, and that the researcher must take the responsibility for assuming that confounding is proprierly addressed in every model tested. However, this is inherently false in many cases. A trivial example comes from the setting of the last simulation above: if a linear relationship is appropriate, the median-split transformation will not provide a test with an adequate control of the Type I error. Conversely, when the dependence should be modeled via a median-split, the natural variables will fail as well. The same can be said for very well known transformations such as log and square-root functions. These considerations shed a light on the implicit complexity of a multiverse analysis. A significant test must be interpreted as a significant relationship between the predictor of interest X and the response Y, given the confounders Z of the model. And the significant result may be due to a real relationship between X and Y or a poor modeling of Z. It is the responsibility of the researcher to carefully consider both the possibilities.

Let's go back to the application in Section 5. The comparison `LockDown - Pre` shows no significant result, therefore no (false) claims can be made. More interestingly, the comparison `Post - LockDown` has 29 significant – i.e. multiplicity corrected – tests. Let's now focus on this comparison. By exploring the results, we can see that most of the significant ones are due to models where the variable `Age` is not transformed (i.e. 27/29), while when the age is modeled by a B-spline transformation, the test becomes not significant in most of the cases (see Table 6). Such a result should cast doubt on the robustness of the results. Most likely, the significant results are due to inadequate modeling of the relationship between `Age` and the response variable which, in turn induces a spurious correlation between the contrast under test and the response variable. In our opinion, therefore, there is not enough evidence to support the claim that the willingness to

Predictor of `Age`	**p-adjusted** > 0.05	**p-adjusted** ≤ 0.05
Age (linear)	0	27
Age with 3 basis of B-splines	27	0
Age with 4 basis of B-splines	25	2

TABLE 6.

Number of models with significant difference `Post - Lock` for different transformations of the variable `Age` in the PIMA of the COVID-19 vaccine hesitancy dataset.

get vaccinated Y has changed between the `Pre` lock-down and the `LockDown` period. This highlights the importance of the multiplicity correction in PIMA to obtain a better understanding of the analysis and results. Particular patterns could suggest, for instance, that significant specifications correspond only to certain modelling choices; the researcher should then evaluate whether these particular choices are indeed plausible or, on the contrary, should be discarded.

Robust analysis is still possible. Despite the challenges pointed out in this discussion, we claim that robust results can still be obtained. Consider the comparison `Post - Pre`, where all comparisons turn out to yield significant effects. If we can assume that "at least one" among all tested models deals properly with confounders, we are allowed to claim that there is a significant difference between `Post` and `Pre` – even though we cannot claim which model is the more appropriate one. This result directly follows from Berger's general results on intersection-union tests (Berger, 1982). Thus, to control the relevant Type I error probability it is only necessary to test each one of the coefficients at the α level.

To conclude, we hope that our proposed inferential framework for multiverse analysis will allow researchers to learn as much as possible from the multiverse analyses they perform. Our extension to generalized linear models allows researchers designing a wider to move beyond a purely descriptive interpretation of a multiverse analysis and allows researchers to test whether the null hypothesis can be statistically rejected in any or a subset of models. The strong control for multiplicity in PIMA provides researchers with a statistical tool that allows them to claim that the null hypothesis can be rejected for each specification that shows a significant effect, with the comfort of knowing that they are not p-hacking. We underline that in scenarios with a large

number of specifications, the correction for multiple testing is essential to understand for which specifications there is a statistically significant relationship and so draw more informative inferences from the data. PIMA makes it possible for researchers who feel that they can not a-priori specify a single analysis approach to efficiently test a plausible set of models and still draw reliable inferences.

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley, New York.

Begg, C. B. and Berlin, J. A. (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 151(3):419–445.

Benjamini, Y. (2020). Selective Inference: The Silent Killer of Replicability. *Harvard Data Science Review*, 2(4). https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Caserotti, M., Girardi, P., Rubaltelli, E., Tasso, A., Lotto, L., and Gavaruzzi, T. (2021). Associations of covid-19 risk perception with vaccine hesitancy over time for italian residents. *Social Science & Medicine*, 272:113688.

De Santis, R., Goeman, J. J., Hemerik, J., and Finos, L. (2022). Inference in generalized linear models with robustness to misspecified variances.

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., and Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS one*, 3(8):e3081.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904.

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *The Annals of Statistics*, 27(1):274–289.

Finos, L. (2022). *jointest: Multivariate testing trought joint sign-flip scores (Hemerik, Goeman and Finos (2020) ¡doi:10.1111/rssb.12369¿).* R package version 1.2.0.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Flachaire, E. (1999). A better way to bootstrap pairs. *Economics Letters*, 64(3):257–262.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

Frey, R., Richter, D., Schupp, J., Hertwig, R., and Mata, R. (2021). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*, 120(2):538.

Gelman, A. and Loken, E. (2014). The statistical crisis in science data-dependent analysis–a "garden of forking paths" – explains why many statistically significant comparisons don't hold up. *American scientist*, 102(6):460.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *J. Am. Statist. Ass.*, 101(476):1408–1417.

Goeman, J. J., Hemerik, J., and Solari, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Statist.*, 49(2):1218–1238.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statist. Sci.*, 26(4):584–597.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological bulletin*, 82(1):1.

Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5):1158–1177.

Hemerik, J. and Goeman, J. J. (2018). Exact testing with random permutations. *TEST*, 27:811–825.

Hemerik, J., Goeman, J. J., and Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):841–864.

Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P., and Boulesteix, A.-L. (2020). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology*, 50(1):266–278.

Liptak, T. (1958). On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 3:1971–1977.

Liu, Y., Kale, A., Althoff, T., and Heer, J. (2020). Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1753–1763.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Mirman, J. H., Murray, A. L., Mirman, D., and Adams, S. A. (2021). Advancing our understanding of cognitive development and motor vehicle crash risk: a multiverse representation analysis. *Cortex*, 138:90–100.

Modecki, K. L., Low-Choy, S., Uink, B. N., Vernon, L., Correia, H., and Andrews, K. (2020). Tuning into the real effect of smartphone use on parenting: a multiverse analysis. *Journal of Child Psychology and Psychiatry*, 61(8):855–865.

Nosek, B. A. and Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, 45(3):137–141.

Open Science Collaboration (2015). *Estimating the reproducibility of psychological science*, volume 349. American Association for the Advancement of Science.

Pesarin, F. (2001). *Multivariate Permutation Tests: with Applications in Biostatistics*. Wiley, New York.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramdas, A., Barber, R. F., Candès, E. J., and Tibshirani, R. J. (2023). Permutation tests using arbitrary permutation distributions. *Sankhya A*, pages 1–22.

Rijnhart, J. J., Twisk, J. W., Deeg, D. J., and Heymans, M. W. (2021). Assessing the robustness of mediation analysis results using multiverse analysis. *Prevention Science*, pages 1–11.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *The Annals of Statistics*, 8(6):1342–1347.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54(285):30–34.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Vesely, A., Finos, L., and Goeman, J. J. (2023). Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85(3).

Vesely, A., Goeman, J. J., and Finos, L. (2022). Resampling-based multisplit inference for high-dimensional regression. https://arxiv.org/abs/2205.12563.

Wessel, I., Albers, C. J., Zandstra, A. R. E., and Heininga, V. E. (2020). A multiverse analysis of early attempts to replicate memory suppression with the think/no-think task. *Memory*, 28(7):870–887.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.