Testing ANOVA Replications by Means of the Prior Predictive p-Value

Mariëlle Zondervan-Zwijnenburg

Utrecht University

Rens van de Schoot

Utrecht University

Herbert Hoijtink

Utrecht University

Author Note

In press at Meta-Psychology. See `osf.io/bm4hx/`.

Correspondence concerning this article should be addressed to Mariëlle Zondervan-Zwijnenburg, Department of Methods and Statistics, Utrecht University, Padualaan 14, 3584CH Utrecht. E-mail: M.A.J.Zwijnenburg@uu.nl

Abstract

In the current study, we introduce the prior predictive p-value as a method to test replication of an analysis of variance (ANOVA). The prior predictive p-value is based on the prior predictive distribution. If we use the original study to compose the prior distribution, then the prior predictive distribution contains datasets that are expected given the original results.

To determine whether the new data resulting from a replication study deviate from the data in the prior predictive distribution, we need to calculate a test statistic for each dataset. We propose to use $\bar{F}$, which measures to what degree the results of a dataset deviate from an inequality constrained hypothesis capturing the relevant features of the original study: H_{RF}. The inequality constraints in H_{RF} are based on the findings of the original study and can concern, for example, the ordering of means and interaction effects. The prior predictive p-value consequently tests to what degree the new data deviates from predicted data given the original results, considering the findings of the original study.

We explain the calculation of the prior predictive p-value step by step, elaborate on the topic of power, and illustrate the method with examples. The replication test and its integrated power and sample size calculator are made available in an R-package and an online interactive application. As such, the current study supports researchers that want to adhere to the call for replication studies in the field of psychology.

Keywords: ANOVA, comparison of means, power analysis, prior predictive p-value, replication study

Testing ANOVA Replications by Means of the Prior Predictive p-Value

Introduction

New studies conducted to replicate earlier original studies are often referred to as replication studies. After the latest "crisis in confidence" in the field of psychology, the call to conduct replication studies is stronger than ever (Anderson & Maxwell, 2016; Asendorpf et al., 2013; Cumming, 2014; Earp & Trafimow, 2015; Ledgerwood, 2014; Open Science Collaboration, 2012, 2015; Pashler & Wagenmakers, 2012; Schmidt, 2009; Verhagen & Wagenmakers, 2014), and large replication projects such as the Reproducibility Project Psychology (Open Science Collaboration, 2015), Reproducibility Project: Cancer Biology (RP:CB) (Errington et al., 2019), and Many Labs projects (Ebersole et al., 2016; Klein et al., 2014, 2018) have been launched. As a result, methodology on conducting replication studies has received increasing attention (see for example Anderson & Maxwell, 2016; Asendorpf et al., 2013; Brandt et al., 2014; Schmidt, 2009). There is, however, no standard methodology to determine whether a replication is successful or not (Open Science Collaboration, 2015).

The results of an original study are replicated when a new study corroborates the original findings. A common and intuitive method to assess whether a result is replicated is 'vote-counting'. Vote-counting is assessing whether the new effect is statistically significant and in the same direction as the significant effect in the original study (Anderson & Maxwell, 2016; Simonsohn, 2015). Vote-counting, however, has serious shortcomings. First of all, it is a dichotomous evaluation that does not take into account the magnitude of differences between effect-sizes of the original and new study (Asendorpf et al., 2013; Simonsohn, 2015). Secondly, each of the effect sizes being significant does not imply that both effect sizes are the same, nor does one significant effect and one non-significant effect imply that both effects are different (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Stated otherwise, vote-counting does not formally test whether a result is replicated (Anderson & Maxwell, 2016; Verhagen & Wagenmakers, 2014). Thirdly, underpowered replication studies are less likely to replicate significance, which can lead to misleading conclusions

(Asendorpf et al., 2013; Cumming, 2008; Hedges & Olkin, 1980; Simonsohn, 2015).

In the current study, we address the following replication research question: "Does the new study fail to replicate relevant features of the original study?". For example, the result of an original ANOVA study is: Group A > Group B > Group C. The finding can be: "Group A performs better than group B, which performs better than group C"; "Group A performs better than group B and C"; or "Group A and B perform better than group C". The 'relevant features' subordinate to the replication test always have to be in line with the original result (i.e., Group A > Group B > Group C) for the test to function properly. If the purpose of the replication test is to put the proclaimed theory by the original to the test, then the claims of the original study determine the exact relevant features to be evaluated. However, if there is reason to test another feature, it is possible to let the relevant features deviate from the claims in the original study. The relevant features of original studies will be captured in the form of an informative hypothesis (Hoijtink, 2012), which is specified using inequality constraints among the means of the ANOVA model. We propose to evaluate the replication of these hypotheses with the prior predictive p-value (Box, 1980).

The prior predictive p-value was not introduced to test replication. It was originally presented as a method to test whether the current data is unexpected given prior expectations concerning the parameter values of a statistical model. A disadvantage of the prior predictive check to test model fit is that it is leaves undetermined whether the prior expectations about the parameter values or the model assumptions are incorrect. Hence, as a model test the prior predictive check has been replaced by the posterior predictive check (Gelman, Meng, & Stern, 1996), which does not make prior assumptions about expected parameter values, but instead uses the posterior results given the current data.

With respect to testing replication, however, the prior predictive check is a good method for three reasons. First, instead of non-empirical prior expectations, we use the posterior distribution of the model parameters given the original data as the prior distribution. Consequently, we have a well-founded and clear-cut prior. Second, the

prior predictive check uses a distribution of datasets (i.e., the prior predictive distribution) that are expected given the prior (i.e., the posterior of the original study). In this manner, the prior predictive distribution takes into account that results in a new dataset - resulting from a replication study - may deviate from the original results because of random variation instead of meaningful differences. According to our definition, a study replicates if the new dataset is drawn from the same population as the original dataset. Third, the prior predictive check uses a 'relevant checking function' for which we propose $\bar{F}$ (Silvapulle & Sen, 2005, p. 38-39). The statistic $\bar{F}$ captures the deviance from a constrained hypothesis that we base on the findings of the original study. As a result, we can check whether the new study significantly fails to replicate relevant features of the original study, while taking variation into account.

Table 1 shows how our research question and proposed method relate to other replication research questions and associated methods that have been proposed. Our method addresses a question similar to that in Anderson and Maxwell (2016); Verhagen and Wagenmakers (2014); Harms (2018); Ly, Etz, Marsman, and Wagenmakers (2018) and Patil, Peng, and Leek (2016), but now enables researchers to evaluate the replication of relevant features of an original ANOVA study. The bottom panel of Table 1 shows other replication research questions that will not be pursued in this paper. The reader interested in these questions, should consult the given references.

The goal of this paper is to introduce the prior predictive p-value as a method to test replication of relevant features of original ANOVA studies. In the first section, we provide a step by step introduction of the prior predictive p-value as included in the `ANOVAreplication` R-package and the online interactive application (see `osf.io/6h8x3`). In the second section, we discuss the statistical power of the prior predictive p-value. In the third section, we explain how to use and interpret the prior predictive p-value by means of a workflow. In the fourth section, we use several studies from the Reproducibility Project Psychology (Open Science Collaboration, 2012) to demonstrate the use of the prior predictive p-value. The paper ends with a discussion and conclusion section.

Table 1

Replication Research Questions and Methods to Address Them

Replication Research Question Current Study and Similar Questions	Method	Setting	Reference
Does the new study fail to replicate relevant features of the original study?	Prior predictive p-value	t-test, ANOVA	Current study
Does the new study fail to replicate the effect size of the original study?	Confidence interval for difference in effect sizes	t-test, correlation	Anderson and Maxwell (2016)
Does the new study replicate the effect size of the original study?	Prediction interval	correlation	Patil et al. (2016)
	Equivalence test	t-test	Anderson and Maxwell (2016)
	Bayes factor	t-test	Verhagen and Wagenmakers (2014)
	Bayes factor	ANOVA	Harms (2018)
	Bayes factor	BF models[a]	Ly et al. (2018)
Other Replication Research Questions	**Method**	**Setting**	**Reference**
Is the effect present or absent in the replication study?	Bayes factor	t-test, correlation[b]	Marsman et al. (2017)
Is Cohen's d in the population of a detectable size?	Telescope test	t-test[c]	Simonsohn (2015)
Is the original effect size extreme in comparison to the new study?	Confidence interval for difference in effect sizes	t-test, correlation	Open Science Collaboration (2015)
What is Cohen's d in the population?	Confidence interval for average effect size	t-test	Anderson and Maxwell (2016)
What is the effect size (corrected for publication bias) in the population?	Hybrid meta-analysis	t-test	Van Aert and Van Assen (2017)

[a]All models for which a Bayes factor can be computed.

[b]The reconceptualization by Ly et al. (2018) generalizes to most common experimental designs.

[c]The telescope test is explained in the t-test setting, but applicable to any model for which a power analysis can be conducted.

Prior Predictive p-Value

The evaluation of the replication of an ANOVA study by means of the prior predictive p-value (Box, 1980) consists of three steps that will be explained below.

Step 1: Prior Predictive Distribution of the Data

The ANOVA model is given by:

$$y_{ijd} = \mu_{jd} + \epsilon_{ijd} \tag{1}$$
$$\epsilon_{ijd} \sim \mathcal{N}(0, \sigma_d^2),$$

where y_{ijd} is observation $i = 1, ..., n_{jd}$ in group $j = 1, ..., J$ for dataset $d \in \{o, r, \text{sim}\}$, where o denotes the original data, r denotes the new data, and sim denotes simulated data, the latter will be introduced towards the end of this section. Furthermore, μ_{jd} is the mean of group j in dataset d, ϵ_{ijd} is the error term, and σ_d^2 is the pooled variance over all J groups.

The original ANOVA results can be summarized in the posterior distribution of the parameters: $g(\boldsymbol{\mu}_o, \sigma_o^2|\boldsymbol{y}_o)$, where $\boldsymbol{\mu}_o = [\mu_{1o}, ..., \mu_{Jo}]$ and $\boldsymbol{y}_o$ includes all observations y_{ijo}:

$$g(\boldsymbol{\mu}_o, \sigma_o^2|\boldsymbol{y}_o) \propto f(\boldsymbol{y}_o|\boldsymbol{\mu}_o, \sigma_o^2)h(\boldsymbol{\mu}_o, \sigma_o^2), \tag{2}$$

where the density of the data

$$f(\boldsymbol{y}_o|\boldsymbol{\mu}_o, \sigma_o^2) = \prod_{j=1}^{J}\prod_{i=1}^{n_{jo}} \frac{1}{\sqrt{2\pi}\sigma_o} e^{\frac{-(y_{ijo}-\mu_{jo})^2}{2\sigma_o^2}} \tag{3}$$

and the standard prior distribution,

$$h(\boldsymbol{\mu}_o, \sigma_o^2) \propto \frac{1}{\sigma_o^2}, \tag{4}$$

that is, a uniform prior on the means and Jeffrey's prior on the variance. The prior distribution for the analysis of the original data is uninformative, that is, the posterior distribution is completely determined by the original data in order to match the results of the original study. If the original study used a Bayesian analysis, the priors should match those of the original study in order to reproduce the original study results. Given

the observed original results, the prior distribution for future parameters $h(\boldsymbol{\mu}_r, \sigma_r^2) = h(\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) = g(\boldsymbol{\mu}_o, \sigma_o^2|\boldsymbol{y}_o)$. With the prior predictive p-value, we then test H_0: $\boldsymbol{\mu}_r, \sigma_r^2 \sim h(\boldsymbol{\mu}_r, \sigma_r^2)$. H_0 states that $\boldsymbol{\mu}_r, \sigma_r^2$ follow the distribution of the prior for $\boldsymbol{\mu}_r, \sigma_r^2$. Loosely formulated, H_0 states that the parameters in the new data are in line with our expectations given the original results.

To test H_0, we obtain datasets that are to be expected given the original data. Using this prior we simulate data $\boldsymbol{y}_{\text{sim}}$ that are to be expected given the results of the original study:

$$f(\boldsymbol{y}_{\text{sim}}) = \int f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) h(\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) d\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2), \tag{5}$$

where $f(\boldsymbol{y}_{\text{sim}})$ is the prior predictive distribution of the data. Note that $f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2)$ is the counterpart of Equation 3 for dataset sim instead of o. Datasets $\boldsymbol{y}_{\text{sim}}^t$ for $t = 1, ..., T$, where T denotes the number of samples from the prior predictive distribution, are obtained by sampling $\boldsymbol{\mu}_{\text{sim}}^t, \sigma_{\text{sim}}^t$ from $h(\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}) = g(\boldsymbol{\mu}_o, \sigma_o|\boldsymbol{y}_o)$, and subsequently simulating $\boldsymbol{y}_{\text{sim}}^t$ from $f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{\mu}_{\text{sim}}^t, \sigma_{\text{sim}}^t)$ (cf. Equation 3). Datasets $\boldsymbol{y}_{\text{sim}}^t$ have sample sizes $n_{1r}, ..., n_{Jr}$, because the predicted data needs to be compared to the new data $\boldsymbol{y}_r$ that has sample sizes $n_{1r}, ..., n_{Jr}$.

The steps in the following sections elaborate how new data $\boldsymbol{y}_r$ can be compared to the T data matrices sampled from $f(\boldsymbol{y}_{\text{sim}})$ that are to be expected given H_0 using a test-statistic that evaluates relevant features of the original data.

Step 2: Test Statistic Evaluating Relevant Features

We propose to use $\bar{F}$ (Silvapulle & Sen, 2005, p. 38-39) as a test-statistic to evaluate how much the predicted data and the observed data deviate from an inequality constrained hypothesis capturing the relevant features of the original study H_{RF}:

$$\bar{F}_{\boldsymbol{y}_d} = \frac{\text{RSS}_{d,H_{\text{RF}}} - \text{RSS}_{d,H_u}}{S_d^2}, \tag{6}$$

where RSS_{d,H_u} denotes the residual sum of squares in dataset $d \in \{r, \text{sim}\}$ for the unrestricted hypothesis H_u: $\mu_{1d}, ..., \mu_{Jd}$,

$$\text{RSS}_{d,H_u} = \sum_{ij} (y_{ijd} - \bar{y}_{jd})^2, \tag{7}$$

where $\bar{y}_{jd}$ denotes the mean for group j in dataset d. S_d^2 denotes the mean squared error,

$$S_d^2 = \frac{\text{RSS}_{d,H_u}}{N - J}, \tag{8}$$

where $N = \sum_{j=1}^{J} n_{jr}$, and

$$\text{RSS}_{d,H_{\text{RF}}} = \sum_{ij} (y_{ijd} - \tilde{\mu}_{jd})^2, \tag{9}$$

where

$$\tilde{\boldsymbol{\mu}}_d = [\tilde{\mu}_{jd}, ..., \tilde{\mu}_{Jd}] = \underset{\tilde{\boldsymbol{\mu}}_d \in H_{\text{RF}}}{\text{argmin}} \sum_{ij} (y_{ijd} - \mu_{jd})^2. \tag{10}$$

$\tilde{\boldsymbol{\mu}}_d$ thus contains the set of parameter estimates that minimize the residual sum of squares for $\boldsymbol{y}_d$ under the constraints imposed by H_{RF}. $\bar{F}_{\boldsymbol{y}_d}$ is the scaled difference between the residual sum of squares under the constraints imposed by H_{RF} and the residual sum of squares for $\boldsymbol{y}_d$ under H_u. As H_u is unrestricted, $\bar{F}_{\boldsymbol{y}_d}$ quantifies the misfit of $\boldsymbol{y}_d$ with H_{RF}.

The hypothesis capturing the relevant features of the original data, H_{RF}, is of the form $\boldsymbol{R}\boldsymbol{\mu}_d > 0$, where $\boldsymbol{R}$ is a $K \times J$ restriction matrix, J denotes the number of groups in the ANOVA study, and K the number of restrictions in H_{RF}, while $\boldsymbol{\mu}_d$ is the mean vector of length J.

Examples of constraints that can be applied under $\boldsymbol{R}\boldsymbol{\mu}_r > 0$ are:

- Simple order constraints: $\mu_{jd} > \mu_{j'd}$, or $\mu_{jd} < \mu_{j'd}$ for a pair j, j'.
- Interaction effects: (μ_{ABd} - $\mu_{AB'd}$) > ($\mu_{A'Bd}$ - $\mu_{A'B'd}$), for a 2×2 contingency table.

The constraints in H_{RF} should be based on the findings of the original study, which implies and requires that H_{RF} is always in agreement with the results of the original study (i.e., $\bar{F}_{\boldsymbol{y}_o} = 0$). The results of the original study alone are usually not enough to determine which H_{RF} is to be evaluated. For example, an original study shows that $\bar{y}_{1o} < \bar{y}_{2o} < \bar{y}_{3o}$. This finding may lead to H_{RF}: $\mu_{1d} < \mu_{2d} < \mu_{3d}$, but also to H_{RF}: $(\mu_{1d}, \mu_{2d}) < \mu_{3d}$ or H_{RF}: $\mu_{1d} < (\mu_{2d}, \mu_{3d})$. Which exact features should be covered in H_{RF} can be guided by the conclusions of the original study. For example, if in the

original study it is concluded that a treatment condition leads to better outcomes than two control conditions, the most logical specification of the relevant features is H_{RF}: $(\mu_{\text{controlA}d}, \mu_{\text{controlB}d}) < \mu_{\text{Treatment}d}$. Alternatively, if in the original study it is concluded that treatment A is better than treatment B, which is better than the control condition, a logical relevant feature hypothesis would be: H_{RF}: $\mu_{\text{TreatmentA}d} > \mu_{\text{TreatmentB}d} > \mu_{\text{Control}d}$. It may also occur that the researcher conducting the replication test has an interest to evaluate a claim that is not in the original study, but could be made based on its results. In all cases, the researcher conducting the replication test should substantiate the choices made in the formulation of H_{RF} with results from the original study. It is good practice to also pre-register H_{RF}. In the Examples Section, we demonstrate for two studies how the original study is linked to H_{RF}. First, however, we explain how the prior predictive p-value is calculated.

Step 3: p-value

The third and final step is to compute the prior predictive p-value. When we calculate $\bar{F}_{\boldsymbol{y}^t_{\text{sim}}}$ for each dataset $\boldsymbol{y}^t_{\text{sim}}$ obtained in Step 1 with respect to $\bar{F}$ as defined in Step 2, a sampling-based representation of the prior predictive distribution of the test statistic $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is obtained. Consequently,

$$p = P(\bar{F}_{\boldsymbol{y}_{\text{sim}}} \geq \bar{F}_{\boldsymbol{y}_r} | H_0) = \frac{1}{T} \sum_{t=1}^{T} I(\bar{F}_{\boldsymbol{y}^t_{\text{sim}}} \geq \bar{F}_{\boldsymbol{y}_r}), \tag{11}$$

where H_0 denotes "Replication", that is: H_0: $\boldsymbol{\mu}_r, \sigma_r^2 \sim h(\boldsymbol{\mu}_r, \sigma_r^2)$. Furthermore, I is an indicator function that takes on the value 1 if the argument is true and 0 otherwise.

As illustrated in Figure 1, the prior predictive p-value indicates how exceptional the observed statistic for the new data, $\bar{F}_{\boldsymbol{y}_r}$, is compared to its prior predictive distribution $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$. The shaded area on the right side of $\bar{F}_{\boldsymbol{y}_r}$ is $P(\bar{F}_{\boldsymbol{y}_{\text{sim}}} \geq \bar{F}_{\boldsymbol{y}_r} | H_0)$, that is, the prior predictive p-value. If the prior predictive p-value is significant, we reject replication of the relevant features of the original study by the new data. Note that the focus is on rejecting replication of the original results and not on rejecting H_{RF}

in itself for the new study.[1]

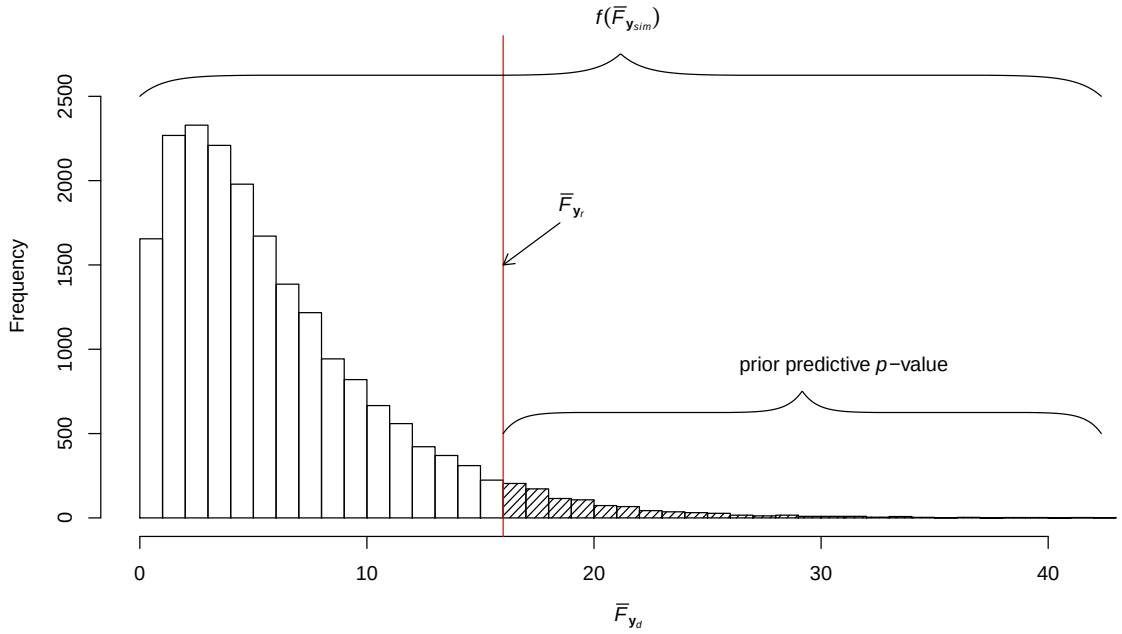


Figure 1. An illustration of the prior predictive p-value.

Uniformity. To determine the significance of a p-value by comparing it to some preselected value α, the p-value needs to be uniformly distributed if H_0 is true. Only when the p-value is uniform, α is equal to the nominal Type I error. We will demonstrate that this is true for the prior predictive p-value if $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is continuous, and it is true up to some α_0 if $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is discrete.

A p-value is uniform if:

$$P(p \leq \alpha | H_0) \leq \alpha \text{ for all } \alpha \in [0, 1], \tag{12}$$

where p denotes a p-value from $f(p|H_0)$, that is, the null-distribution of the p-values.

The following three steps proof that Equation 12 holds for the prior predictive p-value when $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is continuous:

1. $P(p < \alpha | H_0) = P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}}, 1-\alpha} | H_0)$, where $\bar{F}_{\boldsymbol{y}_r}$ is the test-statistic rendering p

[1] To test H_{RF} we recommend Hoijtink, Mulder, van Lissa, and Gu (2019); Vanbrabant, Van de Schoot, and Rosseel (2015).

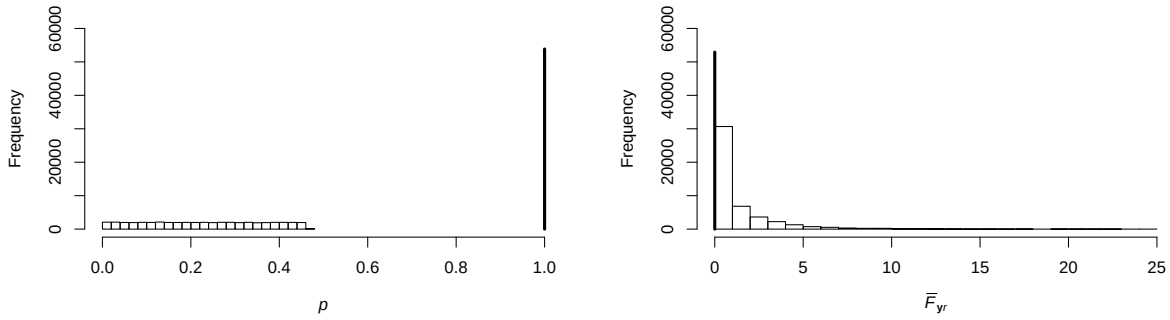
via $p = P(\bar{F}_{\boldsymbol{y}_{\text{sim}}} > \bar{F}_{\boldsymbol{y}_r}|H_0)$ and $\bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}$ is the 1-αth percentile of the distribution $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}|H_0)$.

2. $P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}|H_0) = \int_{\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}} f(\bar{F}_{\boldsymbol{y}_r}|H_0) d\bar{F}_{\boldsymbol{y}_r}$, where $f(\bar{F}_{\boldsymbol{y}_r}|H_0)$ denotes the distribution of $\bar{F}_{\boldsymbol{y}_r}$ under H_0.

3. For the situations considered in this paper it holds that $f(\bar{F}_{\boldsymbol{y}_r}|H_0) = f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$, therefore $\int_{\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}} f(\bar{F}_{\boldsymbol{y}_r}|H_0) d\bar{F}_{\boldsymbol{y}_r} = \int_{\bar{F}_{\boldsymbol{y}_{\text{sim}}} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}} f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}) d\bar{F}_{\boldsymbol{y}_{\text{sim}}} = \alpha$, which completes the proof.

With constraints of the form $\boldsymbol{R}\boldsymbol{\mu}_r > 0$, however, $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ will often be discrete. When $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is discrete, the prior predictive p-value is not uniform for all $\alpha \in [0, 1]$. For example, let us obtain $g(\boldsymbol{\mu}_o, \sigma_o^2|\boldsymbol{y}_o) = h(\boldsymbol{\mu}_r, \sigma_r^2)$ for an original study with $\bar{y}_{1o} = 1, \bar{y}_{2o} = 2, \bar{y}_{3o} = 3, s_o^2 = 5$, and $n_{jo} = 50$, with $n_{jr} = 50$ and H_{RF}: $\mu_{1r} < \mu_{2r} < \mu_{3r}$. Subsequently, we simulate $\boldsymbol{y}_r^t$ for $t = 1, ..., 100,000$, and calculate the prior predictive p-value for each $\boldsymbol{y}_r^t$. The result is $f(p|H_0)$, which is plotted in Figure 2a. In Figure 2a, we see a thick vertical line that indicates a set of p-values with exactly the same value, namely 1.00. This set of equal p-values results from the fact that $H_{\text{RF}} : \mu_{1r} < \mu_{2r} < \mu_{3r}$ is true for a substantial number of datasets $\boldsymbol{y}_r^t$ causing the associated $\bar{F}_{\boldsymbol{y}_r^t}$ to be exactly equal to 0 and the associated prior predictive p-values to be exactly equal to 1 (see Figure 2b). Generally, however, there exists an α_0 for which $f(p|H_0)$ is uniform (Meng, 1994), since all values in $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ other than 0 will occur in a continuous fashion. Thus, α is uniform for $\alpha \in [0, \alpha_0]$. If the preselected $\alpha < \alpha_0$, α is equal to the nominal Type I error. α_0 can be computed as $1 - P(f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}) = 0)$. For example, $\alpha_0 \geq .05$ if no more than 95% of $\bar{F}_{\boldsymbol{y}_{\text{sim}}}$ is exactly 0. It would be exceptional if more than 95% of $\bar{F}_{\boldsymbol{y}_{\text{sim}}} = 0$, but it could occur with extremely low power in the original study and an unspecific H_{RF}. A visualization of $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ can help to roughly estimate α_0. For the discrete $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ considered here 53% of $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}) = 0$ and $\alpha_0 = .47$ (Figure 2b).

In the next section, we deal with another important property of null hypothesis significance testing methods: Power.



(a) $f(p|H_0)$. (b) $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$.

Figure 2. Uniformity of the prior predictive p-value for H_{RF}: $\mu_{1r} < \mu_{2r} < \mu_{3r}$.

Power

Power is the probability to reject the null hypothesis (of replication) with a preselected α when not the null, but an alternative hypothesis is true. Researchers typically pursue a power of .80. Let us denote power by γ.

$$\begin{aligned} \gamma &= P(p < \alpha | H_a), \\ &= P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}} | H_a), \end{aligned} \tag{13}$$

where H_a is the population under the alternative hypothesis for which replication is to be rejected. Note that any population for which H_0 is not true can qualify to reject replication. The population used is determined by the theoretical context in which the replication test takes place. The population with $\mu_{1a} = ... = \mu_{Ja}$ is a special population that is generally considered to display a non-effect in ANOVA studies. Hence, $\mu_{1a} = ... = \mu_{Ja}$ seems a natural default choice for the population under the alternative hypothesis. As a best guess for μ_{ja} and σ_a^2 in a power analysis, the grand mean $\bar{y}_o$ and variance σ_o^2 of the original study can be used. The population under the alternative hypothesis with $\mu_{1a} = ... = \mu_{Ja}$ is on the edge of H_{RF}: it deviates minimally from H_{RF}, hence, the associated γ will be a lower limit. Power will increase when the population under the alternative hypothesis is more different from H_{RF} than in the population with equal means, for example, when the means are ordered differently.

.

Simulation Study

To illustrate the power of the prior predictive p-value, we conducted a simulation study in which we varied the effect size in the original study f_o, the sample size for the original study n_{jo}, the sample size for the new study n_{jr}, the relevant feature of interest H_{RF}, and the population under the alternative hypothesis H_a as specified in Table 2. For each cell in the simulation study, 10,000 samples were drawn from H_a and power was calculated according to Equation 13.

Table 2

Simulation Sample Statistics for Original Study and Population Values under H_a

$\boldsymbol{y}_o$					H_a				
f_o	$\bar{y}_{1o}$	$\bar{y}_{2o}$	$\bar{y}_{3o}$	s_o^2	f_a	μ_{1a}	μ_{2a}	μ_{3a}	σ_a^2
.10	-0.12	0.00	0.12	1.00	0	0.00	0.00	0.00	1.00
.25	-0.31	0.00	0.31	1.00	.10	0.00	-0.24	0.00	1.00
.40	-0.49	0.00	0.49	1.00					

Note. Effect size f as introduced by Cohen (1988, p. 274-275).

Other simulation factors: $n_{jd} \in 20, 50, 100$. H_{RF1}: $\mu_{1d} < (\mu_{2d}, \mu_{3d})$, H_{RF2}: $\mu_{1d} < \mu_{2d} < \mu_{3d}$.

The results of the simulation study are provided in Table 3. As expected, power generally increases with increasing effect sizes, increasing sample sizes, and increasing deviation between $\boldsymbol{y}_o$ and H_a. There are, however, some exceptions: With small f_o and low n_{jo}, larger n_{jr} only emphasize the noise in the original study more and do not lead to an increase in power. Similarly, a more specific H_{RF} does not always increase power. Given original studies with smaller samples and smaller effect sizes, $h(\boldsymbol{\mu}_r, \sigma_r^2)$ is so uninformative that more specific H_{RF} are only more inaccurate under H_0, and $\bar{F}_{\boldsymbol{y}_r}$ needs to be extremely large to reject the null.

Table 3 also shows that the power on the edge (i.e., the power for H_{a1}) is insufficient for original studies with small and medium effect sizes ($\gamma < .60$ in all cells). With medium f_o, power is only sufficient if the new study originates from a population

Table 3

Power

		$H_{\text{RF1}}\|H_{a1}$			$H_{\text{RF2}}\|H_{a1}$			$H_{\text{RF1}}\|H_{a2}$			$H_{\text{RF2}}\|H_{a2}$		
		n_{jr}			n_{jr}			n_{jr}			n_{jr}		
f_o	n_{jo}	20	50	100	20	50	100	20	50	100	20	50	100
.10	20	.03	.01	.00	.02	.00	.00	.11	.08	.05	.08	.04	.02
.10	50	.08	.06	.03	.06	.04	.02	.21	.31	.36	.19	.22	.28
.10	100	.11	.12	.10	.09	.10	.08	.26	.45	.62	.25	.39	.52
.25	20	.13	.10	.06	.09	.05	.01	.33	.40	.45	.25	.26	.23
.25	50	.25	.32	.37	.20	.26	.26	.48	.73	**.88**	.43	.62	**.81**
.25	100	.30	.46	.57	.29	.44	.55	.54	**.83**	**.96**	.53	.79	**.94**
.40	20	.32	.41	.41	.27	.26	.21	.59	.78	**.89**	.49	.64	.75
.40	50	.49	*.66*	*.67*	.45	.68	***.83***	.74	**.93**	**.98**	.69	**.93**	**.99**
.40	100	.55	*.66*	*.67*	.57	***.83***	***.83***	.77	**.93**	**.97**	.77	**.98**	**.99**

Text in cells with $\gamma \geq .80$ is boldface.

Text in cells with a maximum γ in relation to the specific $H_{RF}|H_a$ is italic.

in which the means are ordered differently (e.g., H_{a2}). For original studies with large effect sizes and group sample sizes in the original studies with at least 50 participants per group, power can be sufficient under H_{a1}. Power levels off, however, for H_{RF1} and H_{RF2} at .67, and .83 respectively. Under $\mu_{1a} = \mu_{2a} = \mu_{3a}$, H_{RF1}: $\mu_{1r} < (\mu_{2r}, \mu_{3r})$ is true in $\frac{1}{3}$ of the situations by chance. Consequently, power cannot exceed $1 - \frac{1}{3} = .67$. For H_{RF2}: $\mu_{1r} < \mu_{2r} < \mu_{3r}$, $\frac{1}{6}$ of the combinations under H_{a1} are in line with replication by chance. Hence, power cannot exceed $1 - \frac{1}{6} = .83$. If we move further from the edge of H_{RF}, as we do with H_{a2}, power increases. Thus, the power of the prior predictive p-value considering an H_{RF} with three or fewer order constraints will almost never be high if the true means are equal, but can be high if there is a different ordering in reality as compared to the one in H_{RF}.

The results demonstrate that imprecise estimates (i.e., large standard errors leading to a low informative prior) in the original study lead to low power, especially on

the edge of H_{RF}. This is as true for the prior predictive p-value as it is for other approaches. For example, in a classical ANOVA study with three groups with 20 participants each, power is <.10, <.40, and <.80 for small, medium, and large effect sizes respectively; a result that was already pointed out in Cohen (1988, p. 313). Zondervan-Zwijnenburg and Rijshouwer (2020) demonstrates the application of different methods to evaluate replication, within the context of small samples. Not a single method is unaffected by small sample sizes. As highlighted by Morey and Lakens (n.d.) and Patil et al. (2016): Replication can only be rejected based on the findings of the original study, and when these findings are highly imprecise due to large standard deviations and small sample sizes, rejecting them is hard or even impossible.

Underpowered original studies may result in non-significant prior predictive p-values that have a high probability of being Type II errors (Morey & Lakens, n.d.). Therefore, only reporting the prior predictive p-value is not enough, the probability of a Type II error (i.e., 1-γ) given the population under the alternative hypothesis should be communicated to the reader as well. The next section elaborates on the computation of power and the required sample size for sufficient statistical power. The Workflow and Examples sections explain how researchers should incorporate prior predictive p-values and power. One of the examples will also demonstrate rejected replication despite low power on the edge of H_0.

Power and Sample Size Determination

As highlighted in the previous sections and in the literature (e.g., Brandt et al., 2014; Simonsohn, 2015), power is an important characteristic of a convincing replication study. It is thus important that researchers can calculate the power of the prior predictive check, and can determine the sample size for a new study such that the replication test has high statistical power. Therefore, the `ANOVAreplication` R-package and the online interactive application (see `osf.io/6h8x3`) include a power and sample size calculator.

Given the vector with group sample sizes in the new study $\boldsymbol{n}_r$, $h(\boldsymbol{\mu}_r, \sigma_r^2)$, H_a, H_{RF},

and α, the power γ is calculated as follows:

1. Following Step 1 and 2 of the prior predictive check, $t = 1, ..., T$ datasets are simulated from $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$, and $\bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}}$ can be calculated.

2. Given $\boldsymbol{\mu}_a$, σ_a, $t = 1, ..., T$ datasets are simulated from $f(\bar{F}_{\boldsymbol{y}_{\text{r}}}|H_a)$ with sample sizes $\boldsymbol{n}_r$. Following Step 2 of the prior predictive check, for each dataset $\bar{F}_{\boldsymbol{y}_r}$ is calculated.

3. $\gamma = P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}}|H_a) =$
$$\frac{1}{T}\sum_{t=1}^{T} I(\bar{F}_{\boldsymbol{y}_r^t} \geq \bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}}),$$

As default choice for $\boldsymbol{\mu}_a$, we recommend to use $\bar{y}_o$ for each group. With this setting, the power is calculated to reject replication in case of equal group means. As default choice for σ_a, we recommend the pooled standard deviation of the original study.

To determine the required sample size to reject replication with sufficient power, we use an iterative procedure. In addition to $h(\boldsymbol{\mu}_r, \sigma_r^2)$, H_a, H_{RF}, α, we use the following information to calculate the required sample size: a target power level $\tilde{\gamma}$; a small margin covering acceptable values around the target power γ_{margin}, because the calculated power may not be exactly equal to the target power; a starting value for the group sample size n_{jr_0}; a maximum number of iterations $Q_{\max}$; and a maximum total sample size for the new study $N_{r_{\max}}$. Our default values are: $\tilde{\gamma} = .825$, $\gamma_{\text{margin}} = .025$, $\alpha = .05$, $n_{jr_0} = 20$, $Q_{\max} = 10$, and $N_{r_{\max}} = 600$.

1. In every iteration q, γ_q is calculated given n_{jr_q}.

2. When $q > 1$, $n_{jr_{q+1}}$ is determined by regressing $\{\gamma_1, ..., \gamma_i\}$ on $\{n_{jr_1}, ..., n_{jr_q}\}$ with a linear or quadratic (only if $q = 3$) function. In case of a linear regression, the linear regression coefficient β_1 is the power increase per subject. Subsequently, $n_{jr_{q+1}} = (\gamma_q - \tilde{\gamma})/\beta_1 + n_{jr_q}$. In case of regression with a quadratic function, $n_{jr_{q+1}}$ is calculated by solving the polynomial: $\tilde{\gamma} = \beta_0 + \beta_1 n_{jr_{q+1}} + \beta_2^2 n_{jr_{q+1}}$.

3. Repeat step (1) and (2) until $\gamma_q \in [\tilde{\gamma} - \gamma_{\text{margin}}, \tilde{\gamma} + \gamma_{\text{margin}}]$ (i.e., power is sufficient), or $\gamma_{q-1} \approx \gamma_q$ (i.e., power does not increase anymore up to two decimal

points), or $n_{jr_{q-1}} = n_{jr_q}$ (i.e, the sample size does not change anymore), or $q = Q_{\max}$, or $\Sigma_{j=1}^{J} n_{jr_q} = N_{\max}$.

Workflow

To clarify the procedure to obtain the prior predictive p-value, the workflow is depicted in Figure 3.

Step 1. The first steps (1a-1c) only require the original study. Step 1a is to derive the relevant feature to be evaluated in the test statistic from the findings of the original study. Next, the population for which replication should be rejected (i.e., H_a) can be defined. What is the ordering of the means in this population and what is the effect size in that ordering? H_a can be a population in which all means are equal, but it does not have to be. Step 1c is to obtain the data of the original study, or reconstruct the data based on reported means, standard deviations and group sample sizes. If the new study is not yet conducted, the second step is to calculate the required sample size per group for the new study to reject replication with sufficient power (i.e., γ).

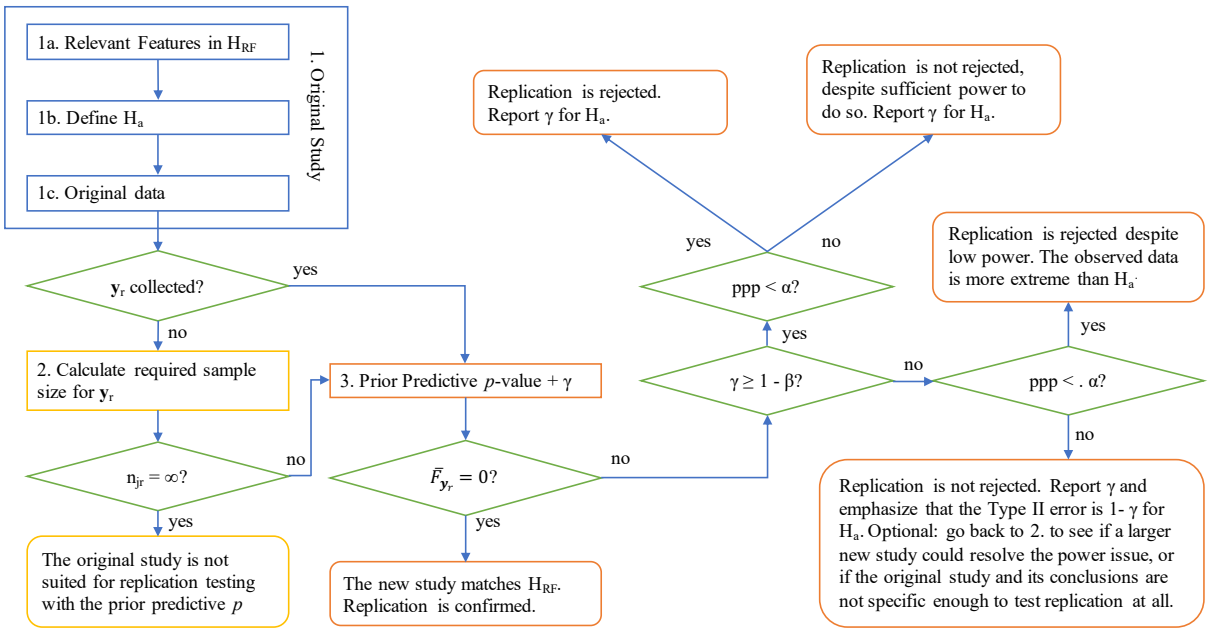


Figure 3. The prior predictive p-value workflow.

Step 2. The sample sizes calculation can be conducted with the `sample.size.calc` function in the `ANOVAreplication` package. If the function cannot

find a (reasonable) group sample size for which γ is sufficient, this implies that the original study is not suited for replication testing with the prior predictive p-value for the specified H_a: its conclusions are too vague (i.e., the standard errors are too wide) to reject replication if H_a is true. There is still a chance that the prior predictive p-value turns out significant, especially if the observed data is more extreme than most samples from H_a, but the researcher should consider whether collecting data with such a low probability of a meaningful result is ethically acceptable.

Step 3. As a third step, the prior predictive p-value can be computed with the function `prior.predictive.check`. The power associated to the sample size of the new study can be calculated with `power.calc`. Note that it is not a post-hoc power analysis, as the definition of H_a is unrelated to the new study. Hence, the power to reject replication for H_a can be insufficient (i.e., larger than 1- the preset Type II error rate β), while the prior predictive p-value is statistically significant, or vice versa. Figure 3 assists in interpreting the resulting p-value, considering the statistical power to reject replication for H_a, unless $\bar{F}_{\boldsymbol{y}_r}$ is exactly 0. If the new study perfectly meets the features of the original study as described in H_{RF}, $\bar{F}_{\boldsymbol{y}_r}$ will be 0 and the prior predictive p-value 1.00. In such a case, we confirm replication of the relevant features in the original study as captured in H_{RF}, irrespective of power. Theoretically it is possible that $\bar{F}_{\boldsymbol{y}_r} = 0$, while the new study is an extreme sample from a population in which H_{RF} is not true. That, however, is not under consideration here, as our question was whether the observed new study replicates, or fails to replicate, relevant features of the original study.

In case of a non-significant result in combination with low power, the researcher should emphasize the probability that not rejecting replication is a Type II error, and it is advised to conduct a replication study with larger n_{jr}. The required sample size per group can again be calculated with the `sample.size.calc` function in the `ANOVAreplication` package. If the required n_{jr} is excessive given H_a, it may be an inevitable conclusion that the original study is not suited for replication testing by means of the prior predictive p-value. If replication is rejected despite low power, it

implies that the observed new dataset deviates more from H_{RF} than most datasets under H_a. With sufficient statistical power, it is still informative to notify the reader of the achieved power and/or the probability of a Type II error given the population under H_a.

Examples

To illustrate the use of the prior predictive check to assess whether relevant ANOVA features are replicated, we two selected replication studies that were part of the Reproducibility Project Psychology initiated by the Open Science Collaboration (2012, 2015). All calculations can be performed with the `ANOVAreplication` R-package (Zondervan-Zwijnenburg, 2018).

The first study is Fischer, Greitemeyer, and Frey (2008), who studied the impact of self-regulation resources on confirmatory information processing. According to the theory, people who have low self-regulation resources (i.e., depleted participants) will prefer information that matches their initial standpoint. An ego-threat condition was added, because the literature proposes that ego-threat affects decision relevant information processing, although the direction of this effect is not clear. To determine which relevant feature of the results (see Table 4) should be tested for replication, we follow the original findings: "Planned contrasts revealed that the confirmatory information processing tendencies of participants with reduced self-regulation resources [...] were stronger than those of nondepleted [...] and ego threatened participants [...]" Fischer et al. (2008, p. 387). This translates to: H_{RF}: $\mu_{\text{low self-regulation},r} > (\mu_{\text{high self-regulation},r}, \mu_{\text{ego-threatened},r})$ (Workflow Step 1a). We want to reject replication when all means in the population are equal. That is: H_a: $\mu_{\text{low self-regulation},r} = (\mu_{\text{high self-regulation},r} = \mu_{\text{ego-threatened},r})$ (Workflow Step 1b). We simulate the original data based on the means, standard deviations and sample sizes reported in Fischer et al. (2008) (Workflow Step 1c). As the replication study is already conducted by Galliani (2015) (see Table 4 for results), we do not calculate the required sample size to test replication (Workflow Step 2), and proceed to calculate the prior predictive p-value and

the power of the replication test (Workflow Step 3). The resulting prior predictive p-value was .003 with $\gamma = .66$, indicating that we reject replication, despite limited power. The ordering in the new data by Galliani (2015) results in an extreme $\bar{F}$ score compared to the predicted data. Figure 4 illustrates this conclusion: Over 90% of the predicted data scores perfectly in line with H_{RF}, but the new study by Galliani (2015) deviates from H_{RF} and scores in the extreme 0.3% of the predicted data. The replication of the original study conclusions is thus rejected.

Table 4

Descriptive Statistics for Confirmatory Information Processing from the Original Study: Fischer et al. (2008), and the New Study: Galliani (2015)

	Low self-regulation		High self-regulation		Ego-threatened	
Study	n	M (SD)	n	M (SD)	n	M (SD)
Original	28[a]	0.36 (1.08)	28[a]	-0.19 (0.53)	28[a]	-0.18 (0.81)
New	48	-0.07 (0.45)	47	-0.05 (0.47)	45	0.13 (0.64)

[a]Only the total sample size of 85 was provided in Fischer et al. (2008).

The second study is Janiszewski and Uy (2008), who studied numerical judgements with five experiments. More specifically, they study the impact of precision of an anchor, and motivation to adjust from the anchor on judgement bias. The group means, standard deviations, and sample sizes of experiment 4a in the original study by Janiszewski and Uy (2008) and the replication study by Chandler (2015) are provided in Table 5. We find that based on these results, Janiszewski and Uy (2008) draw two conclusions. "First, a precise anchor results in less adjustment than a rounded anchor" (p. 126). For experiment 4a, which was replicated by Chandler (2015), this conclusion translates to H_{RF}: $(\mu_{\text{low motivation,round},r} > \mu_{\text{low motivation,precise},r})$ & $(\mu_{\text{high motivation,round},r} > \mu_{\text{high motivation,precise},r})$ (Workflow Step 1a). We want to reject replication when all means in the population are equal. That is: H_a:

$$\mu_{\text{low motivation,round},r} = \mu_{\text{low motivation,precise},r} = \mu_{\text{high motivation,round},r} = \mu_{\text{high motivation,precise},r}$$

(Workflow Step 1b). We simulate the original data based on the means, standard deviations and sample sizes reported in Janiszewski and Uy (2008) (Workflow Step 1c).

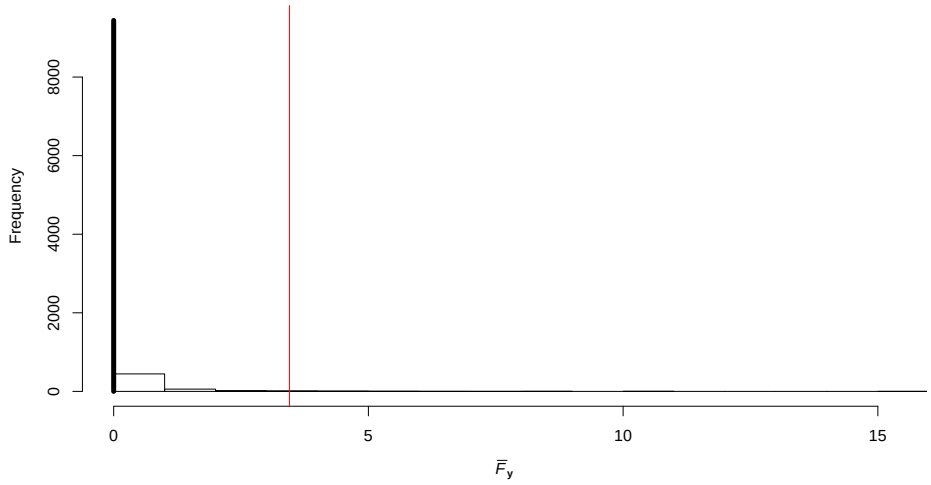


Figure 4. The prior predictive p-value for the replication of Fischer et al. (2008) by Galliani (2015). The histogram bars represent $\bar{F}$ for the predicted data. The thick line on the left represents $\bar{F}$ for the predicted data that are exactly 0 (i.e., over 90% of the total), whereas the red line represents $\bar{F}$ for Galliani (2015).

As the replication study is already conducted by Chandler (2015), we do not calculate the required sample size to test replication (Workflow Step 2), and proceed to calculate the prior predictive p-value and the power of the replication test (Workflow Step 3). The resulting prior predictive p-value is 1.00. The data obtained by Chandler (2015) were perfectly in line with the H_{RF} describing the effect as observed by Janiszewski and Uy (2008). Therefore, we do not have further concerns about the obtained power. Hence, we conclude that the results of Janiszewski and Uy (2008) with respect to H_{RF}: $(\mu_{\text{low motivation,round},r} > \mu_{\text{low motivation,precise},r})$ & $(\mu_{\text{high motivation,round},r} > \mu_{\text{high motivation,precise},r})$ are replicated by Chandler (2015).

The other conclusion that Janiszewski and Uy (2008) draw is about the presence of an interaction effect of adjustment motivation and anchor rounding: "The difference in the amount of adjustment between the rounded- and precise-anchor conditions increased as the motivation to adjust went from low [...] to high" (p. 125). The results and conclusions of Janiszewski and Uy with respect to experiment 4a translate to H_{RF}: $(\mu_{\text{low motivation,round},r} > \mu_{\text{low motivation,precise},r})$ & $(\mu_{\text{high motivation,round},r} >$

Table 5

Z-scores of Participants' Mean Estimates from the Original Study: Janiszewski and Uy (2008), and the New Study: Chandler (2015)

	Low Motivation to Adjust				High Motivation to Adjust			
	Precise Anchor		Rounded Anchor		Precise Anchor		Rounded Anchor	
Study	*n*	*M (SD)*	*n*	*M (SD)*	*n*	*M (SD)*	*n*	*M (SD)*
Original	14	-0.76 (0.17)	15	-0.23 (0.48)	15	-0.04 (0.28)	15	0.98 (0.41)
New	30	-0.35 (0.23)	30	-0.18 (0.37)	30	0.20 (0.34)	30	0.35 (0.44)

$\mu_{\text{high motivation,precise},r}) \;\&\; (\mu_{\text{low motivation,round},r} - \mu_{\text{low motivation,precise},r}) < (\mu_{\text{high motivation,round},r} - \mu_{\text{high motivation,precise},r})$. The prior predictive p-value related to this H_{RF} is .014 with $\gamma = .87$. Thus, we reject replication of the interaction effect.

Discussion & Conclusion

The goal of the current paper was to introduce the prior predictive check as a manner to test replication of ANOVA features. With the prior predictive check researchers can find an answer to the question: "Does the new study fail to replicate relevant features of the original study?" Identifying a non-replication may make us wonder about the representativeness of the original study, the new study, and the comparability of both studies. Or, as stated by Simonsohn (2015, p. 9) "Statistical techniques help us identify situations in which something other than chance has occurred. Human judgment, ingenuity, and expertise are needed to know what has occurred instead."

In the current paper, we discussed the prior predictive p-value for the ANOVA setting. For the ANOVA setting, we explained how to test relevant features of the form $\boldsymbol{R}\boldsymbol{\mu}_r > 0$. Technically, however, the relevant features evaluated by the `ANOVAreplication` R-package, however, can also be of the form $\boldsymbol{R}\boldsymbol{\mu}_r > \boldsymbol{r}$ and $\boldsymbol{S}\boldsymbol{\mu}_r = \boldsymbol{s}$, where $\boldsymbol{r}$ and $\boldsymbol{s}$ are vectors of length K containing the constants in H_{RF}, and $\boldsymbol{S}$ is a $K \times J$ restriction matrix like $\boldsymbol{R}$. Accordingly, minimum (effect size) differences between means can be evaluated and means can be constrained equal to specific values. Even

though constraints of these forms can be evaluated with the R-package and in the online application, they are not emphasized in the current paper because they will less often directly relate to the findings of an original study.

The prior predictive p-value is generalizable to statistical models other than the ANOVA as well. That is, for any model a predictive distribution can be obtained, constrained hypotheses can be constructed, and a test-statistic evaluating the constraints can be calculated. The test as currently provided can already be used for the repeated measures ANOVA by means of contrast weights (see, for example, Furr & Rosenthal, 2003). With contrast weights a score for each participant can be calculated indicating to what degree the participant follows the expected pattern. Subsequently, the replication of relevant features of these contrast scores over groups can be tested. A pre-print introduction to test replication with the prior predictive p-value for structural equation models has been published at `https://psyarxiv.com/uvh5s`.

In the current paper, we introduced the prior predictive p-value as a new tool to quantify replication failure or success to the meta-scientific toolbox. With the prior predictive p-value we test whether the new study significantly deviates from our expectations based on the original study. Other methods to evaluate replication research questions with are included in Table 1 and demonstrated in Zondervan-Zwijnenburg and Rijshouwer (2020). Two features of the prior predictive p-value to test replication stand out. First, the prior predictive p-value makes use of a predictive distribution given the original study results. The new study results are compared to the predicted data. A Bayes factor on the other hand, weighs the evidence for two competing hypotheses in the new study as it actually occurred, but does not take study variation into account. Second, to compare the new study with the predicted data, we consider relevant features of the original study. While most other methods evaluate the replication of a simple effect size, relevant features can be any constraint or set of constraints of the form $\boldsymbol{R}\boldsymbol{\mu}_r > 0$, which seamlessly connects to the research objective of most ANOVA studies. With the `ANOVAreplication` R-package including a vignette as a tutorial, and the interactive application (see `osf.io/6h8x3`), we provide

researchers with an easy to use test for replication of ANOVA features. The availability of the prior predictive p-value to test replications can further promote the trend to conduct more replication studies in the field of psychology.

Author Contributions

MZ and HH were involved in the initial research design. MZ drafted and revised the article in collaboration with HH. MZ developed the interactive application, conducted the simulation studies, and conducted the analyses. RS provided additional feedback, and evaluated the interactive application. All authors approved the final manuscript.

Acknowledgements

We would like to thank Meta-Science editor dr. Felix Schönbrodt, and reviewers dr. Matt Williams and dr. Zoltan Dienes for their helpful feedback on this manuscript.

The first and third author are supported by the Consortium Individual Development (CID), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). The second author is supported by a VIDI grant from the Netherlands Organization for Scientific Research (NWO grant number 452.14.006).

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1-12. doi: 10.1037/met0000051

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. doi: 10.1002/per.1919

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, *143*(4), 383–430. doi: 10.2307/2982063

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. doi: 10.1016/j.jesp.2013.10.005

Chandler, J. (2015). *Replication of Janiszewski & Uy (2008, PS, study 4b).* online. Open Science Framework. Retrieved from `osf.io/aaudl`

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9780203771587

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300. doi: 10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi: 10.1177/0956797613504966

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. doi: 10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many labs 3: Evaluating participant pool

quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. doi: 10.1016/j.jesp.2015.10.012

Errington, T., Tan, F., Lomax, J., Perfito, N., Iorns, E., Gunn, W., & Lehman, C. (2019, November). *Reproducibility project: Cancer biology.* Retrieved from `osf.io/e81xl`

Fischer, P., Greitemeyer, T., & Frey, D. (2008). Self-regulation and selective exposure: The impact of depleted self-regulation resources on confirmatory information processing. *Journal of Personality and Social Psychology*, *94*(3), 382. doi: 10.1037/0022-3514.94.3.382

Furr, R. M., & Rosenthal, R. (2003). Repeated-measures contrasts for "multiple-pattern" hypotheses. *Psychological Methods*, *8*(3), 275-293. doi: 10.1037/1082-989X.8.3.275

Galliani, E. (2015). *Replication report of Fischer, Greitemeyer, and Frey (2008, JPSP, study 2).* online. Retrieved from `osf.io/j8bpa`

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. doi: 10.1198/000313006x152649

Harms, C. (2018). A bayes factor for replications of anova results. *The American Statistician*. doi: 10.1080/00031305.2018.1518787

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*(2), 359. doi: 10.1037/0033-2909.88.2.359

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press. doi: 10.1201/b11158

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*(5), 539–556. doi: 10.1037/met0000201

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of

adjustment. *Psychological Science*, *19*(2), 121–127. doi: 10.1111/j.1467-9280.2008.02057.x

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A'many labs' replication project. *Social Psychology*, *45*(3), 142-152. doi: 10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., ... Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. doi: 10.1177/2515245918810225

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, *9*(3), 275–277. doi: 10.1177/1745691613513470

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication bayes factors from evidence updating. *Behavior Research Methods*, 1-11. doi: 10.3758/s13428-018-1092-x

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'replications of important results in social psychology'. *Royal Society Open Science*, *4*(1), 160426. doi: 10.1098/rsos.160426

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142-1160. doi: 10.1214/aos/1176325622

Morey, R. D., & Lakens, D. (n.d.). *Why most of psychology is statistically unfalsifiable.* doi: 10.5281/zenodo.838685

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. doi: 10.1038/nn.2886

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological*

Science, *7*(6), 657–660. doi: 10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in Psychological Science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi: 10.1177/1745691612465253

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544. doi: 10.1177/1745691616646366

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90-100. doi: 10.1037/a0015108

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Order, inequality, and shape constraints* (Vol. 912). John Wiley & Sons. doi: 10.1002/9781118165614

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559-569. doi: 10.1177/0956797614567341

Van Aert, R. C., & Van Assen, M. A. (2017). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, 1-25. doi: 10.3758/s13428-017-0967-6

Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for ANOVA and regression. *Frontiers in Psychology*, *5*, 1565. doi: 10.3389/fpsyg.2014.01565

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457-1475. doi: 10.1037/a0036731

Zondervan-Zwijnenburg, M. A. J. (2018). ANOVAreplication: Test ANOVA replications by means of the prior predictive p-value [Computer software manual].

Retrieved from `https://CRAN.R-project.org/package=ANOVAreplication` (R package version 1.1.3)

Zondervan-Zwijnenburg, M. A. J., & Rijshouwer, D. (2020). Testing replication with small samples: Applications to ANOVA. In R. van de Schoot & M. Miocevic (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners.* Routledge.