12

TESTING REPLICATION WITH SMALL SAMPLES

Applications to ANOVA

Mariëlle Zondervan-Zwijnenburg

DEPARTMENT OF METHODOLOGY AND STATISTICS, UTRECHT UNIVERSITY, UTRECHT, THE NETHERLANDS

Dominique Rijshouwer

DEPARTMENT OF PSYCHOLOGY, UTRECHT, THE NETHERLANDS

Introduction

Concerns about the replicability of studies were expressed as early as in 1979 by Robert Rosenthal, who believed that future insights would solve this problem. However, the field of psychological science is still struggling to establish replicability, as was clearly shown with the Reproducibility Project: Psychology (RPP; Open Science Collaboration, 2015). Increased awareness of the noisiness of results obtained using small samples is an important step towards improving this situation (Lindsay, 2015). Results obtained with smaller samples are less likely to be replicated than those obtained with larger samples (Cohen, 1962).

One of the difficulties in replicating small sample research is that small samples are particularly sensitive to "researcher degrees of freedom": decisions that researchers make in the design and analysis of the data (Simmons, Nelson, & Simonsohn, 2011). For example, researchers decide to combine categories, exclude scores, add comparisons, add covariates, or transform measures. Unfortunately, modifications are more common if results do not support the hypothesis. For example, the impact of an extreme score will more often be detected and adjusted if it causes a non-significant result as compared to a significant result. With small samples, these decisions can easily affect the significance of results, leading to inflated false-positive rates (Simmons et al., 2011).

Another issue is publication bias: studies with statistically significant results are published more often than studies with non-significant results. Small sample studies are often underpowered, leading to non-significant results and hence a reduced chance to be published. On the other hand, small studies that do find significant effects appear impressive and are more likely to be published.

Thus, researcher degrees of freedom and publication bias can lead to overestimation of effects and an inflated false-positive rate in the literature (Simmons et al., 2011). Small sample findings therefore can easily be spurious, meaning that their replication is of great importance.

Different replication research questions require different methods. Here, we distinguish four main research questions that can be investigated if a new study is conducted to replicate an original study:

1. Is the new effect size similar to the original effect size?
2. Is the new effect size different from the original effect size?
3. Are the new findings different from the original findings?
4. What is the effect size in the population?

Note that questions 1 and 2 differ in where the burden of proof lies. Question 1 looks to *provide support* for the equality of effect sizes, whereas question 2 is aimed at *falsifying* the claim of equality of effect sizes in favor of a conclusion that the effect size was not replicated.

For all four replication research questions we recommend statistical methods and apply them to an empirical example. Note that Anderson and Maxwell (2016) also documented replication research questions and associated methods, although not specifically for small samples. In the current chapter, we adopt several suggestions from Anderson and Maxwell (2016) and add more recent methods. R-code (R Core Team, 2017) for reproducing all chapter results is provided as Supplementary Material available on the Open Science Framework (https://osf.io/am7pr/). We demonstrate the four replication research methods for the replication of Henderson, De Liver, and Gollwitzer (2008) by Lane and Gazerian (2016). First, we introduce the original study by Henderson et al. and its replication by Lane and Gazerian. This is followed by a discussion of the four replication research questions and their associated methods.

Example: original study and its replication

Henderson et al. (2008) conducted a series of experiments showing that people who planned the implementation of a chosen goal (i.e., people with an "implemental mind-set") have stronger attitudes, even towards topics unrelated to their actions. Experiment 5 is the one that was replicated by Lane and Gazerian (2016). It is designed to demonstrate that a focus on information that supports the previously made decision is the reason that attitude strength increases with an implemental mind-set. The experiment included three conditions with 46 participants in total. The first condition was a neutral condition in which participants described things they did on a typical day. The second condition was an implemental one-sided focus condition. Participants in this condition chose a romantic topic to write about and wrote down three reasons for that choice. The third condition was the implemental two-sided focus condition in which participants made their choice and wrote down three reasons for and three reasons against this choice. Afterwards, participants in all conditions answered three questions rating their attitude ambivalence with respect to the issue of making public a list with names of convicted sex offenders (e.g., "I have strong mixed emotions both for and against making the list of convicted sex offenders available to the general public rather than just the police").

The descriptive statistics of the data from the experiment by Henderson et al. (2008) are provided in Table 12.1. The effect of the conditions on attitude ambivalence was significant, using an alpha value of .05, as Henderson et al. report: $F(2,\ 43) = 3.36, p = .044, \eta^2 = .13, \omega^2 = .09, r = .26$. We have added the effect size ω^2, because it is less biased than effect size η^2 for small samples (Okada, 2013). Furthermore, we also computed the effect size r as used in the RPP as an additional effect size measure (see Appendix 3 of the RRP available at osf.io/z7aux). Assuming that all predictors (i.e., the dummy condition variables) contributed equally to the explained variance, r^2 is the explained variance per predictor, and r is the correlation coefficient per predictor. If the conditions did not contribute equally, r^2 and r are the average explained variance or correlation coefficient per condition.

Post hoc comparisons revealed that the implemental mind-set, one-sided group demonstrated significantly lower amounts of ambivalence compared to the implemental mind-set, two-sided, group: $t(28) = 2.45$, $p = .021$, Cohen's $d = .93$, Hedges' $g = .50$. For the t-test we added Hedges' g to correct for an upward bias that Cohen's d shows with small samples. Hedges' g is obtained by multiplying Cohen's d by the correction factor $(1 - \frac{3}{4\text{df}-1})$ (Hedges, 1981). The mean of the neutral mind-set group was in the middle, but it was not significantly higher or lower than the means of other conditions (see descriptive statistics in Table 12.1). Henderson et al. (2008) write: "Critically, the findings showed that it was the evaluatively one-sided analysis of information, rather than simply the act of deciding itself, that fostered a spillover of decreased ambivalence" (pp. 406–407).

Lane and Gazerian (2016) replicated the experiment with 70 participants, but found no significant effect of condition on ambivalence: $F(2,\ 67) = 1.70, p = .191$,

TABLE 12.1 Descriptive statistics for confirmatory information processing from the original study: Henderson et al. (2008), and the new study: Lane and Gazerian (2016)

	Neutral		*One-sided implemental*		*Two-sided implemental*	
	n	*μ (SD)*	*n*	*μ (SD)*	*n*	*μ (SD)*
Original	16	1.23 (1.64)	15	0.16 (1.85)	15	1.82 (1.86)
New	24	-0.38 (1.44)	23	-0.14 (1.66)	23	0.39 (1.25)

$\eta^2 = .05$, $\omega^2 = .02$, $r = .16$ (see also the descriptive statistics in Table 12.1). The post hoc difference test between the one-sided and two-sided implemental mind-set groups was not significant either: $t(44) = 1.24$, $p = .222$, Cohen's $d = .36$, Hedges' $g = .25$. Based on the lack of significance in the new study, Lane and Gazerian conclude that the effect may not replicate.

Four replication methods

Evaluating the significance (and direction) of the effect in the new study and using it as a measure for replication, as was a main method of Lane and Gazerian, is called "vote-counting". Vote-counting, however, does not take into account the magnitude of the differences between effect sizes (Simonsohn, 2015), it is not a statistical test of replication (Anderson & Maxwell, 2016; Verhagen & Wagenmakers, 2014), and it leads to misleading conclusions in underpowered replication studies (Simonsohn, 2015). Thus, vote-counting is a poor method to assess replication. In the following, we discuss four alternative replication research questions and methods.

Question 1: Is the new effect size similar to the original effect size?

A frequentist approach to this replication research question is the equivalence test (e.g., Walker & Nowacki, 2011). This test requires the researcher to specify a region of equivalence for the difference between the original and new effect size. If the confidence interval of the difference between effects falls entirely within this region, the effect sizes are considered equivalent. However, it is difficult to set a region of equivalence that is reasonably limited while at the same time the confidence interval for the difference between effects has a chance to entirely fit within the interval. Therefore, we do not elaborate on the equivalence test and focus instead on Bayesian approaches.

To evaluate whether the new effect size is similar to the original effect size, we can compute a Bayes factor (BF; Jeffreys, 1961); see also Chapter 9 (Klaassen). A BF expresses the shift in belief, relative to our prior belief, after observing the data for two competing hypotheses. A BF of 1 is undecided. BFs smaller than 1 indicate preference for the null hypothesis, whereas BFs larger than 1 favor the alternative hypothesis. The two competing hypotheses in the BF can be operationalized in many ways, but in the replication setting, one of the evaluated hypotheses is often the null effect (i.e., the effect size is zero). To evaluate the current research question, a proper alternative hypothesis is that the effect in the new study is similar to the effect in the original study (Harms, 2018a; Ly, Etz, Marsman, & Wagenmakers, 2018; Verhagen & Wagenmakers, 2014). In this case, the BF evaluates whether the new study is closer to a null effect, or closer to the original effect, where the original effect forms the prior distribution in the BF for the new effect. Verhagen and Wagenmakers (2014) developed this BF for the *t*-test. Harms (2018a) extended the Replication BF to the ANOVA *F*-test and developed the `ReplicationBF` R package to compute it based on the sample sizes and test statistics of the original and new study. For the ANOVA by Henderson et al. (2008) replicated by Lane and Gazerian (2016), we obtain a Replication BF of 0.42[1], which means that the evidence for the null hypothesis of no effect is 2.40 (i.e., 1/0.42) times stronger than the evidence for the alternative hypothesis that the effect is similar to that in the original study. See Figure 12.1 for a visualization by the `ReplicationBF` package. The `R` package also includes the Replication BF for *t*-tests as proposed by Verhagen and Wagenmakers (2014). For the post hoc *t*-test we find a Replication BF of .72, which is again in favor of a null effect. Thus, the Replication BF does not support replication of the omnibus ANOVA effect, nor does it support the replication of the post hoc result that the one-sided mind-set group scores lower on ambivalence than the two-sided mind-set group.

Ly et al. (2018) provided a simple calculation to obtain the Replication BF by Verhagen and Wagenmakers (2014) for all models for which a BF can be obtained: Evidence Updating (EU) Replication

$$\text{BF} = \frac{\text{BF combined data}}{\text{BF original data}} \tag{12.1}$$

This calculation (12.1) assumes, however, that the data are exchangeable (see Chapter 2 for a discussion on exchangeability; Miočević, Levy, & Savord). If the original and new study are not based on the same population, the combined data may demonstrate artificially inflated variances due to different means and standard deviations. To minimize the impact of non-exchangeable datasets, Ly et al. (2018) suggest transforming the data. Here, the grand mean

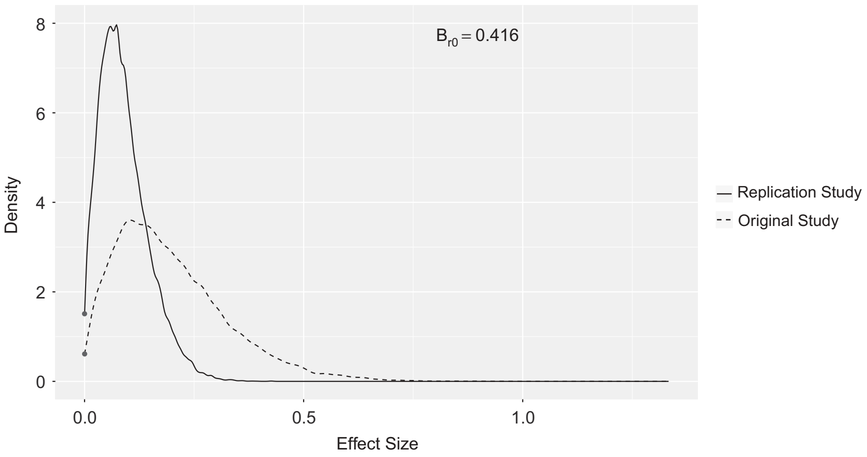

FIGURE 12.1 The Replication BF by Harms (2018a). The original study is the prior for the effect size and the replication study is the posterior based on that prior and the new study. The ratio of the two distributions at 0 on the x-axis is the Replication BF

in Henderson et al. (2008) is actually 1.03 points higher than the grand mean in Lane and Gazerian (2016). To address this issue, we converted the responses to *Z*-scores.

To compute the BFs for the combined and original datasets, we can use the point-and-click software `JASP` (JASP Team, 2018) or the `BayesFactor` package (Morey & Rouder, 2018) in `R`. For both software packages, the BF for the combined data is 1.50, and the BF for the original data is 1.59. Hence, the EU Replication BF = 1.50/1.59 = .94, which favors the ANOVA null hypothesis that the effect is zero. For the post hoc analysis with the alternative hypothesis that the one-sided mind-set group scores lower than the two-sided mind-set group, the BF for the combined data is 6.66 (see Figure 12.2 for the accompanying `JASP` plot) and the BF for the original data is 5.81. Hence, the EU Replication BF = 6.66/5.81 = 1.15 for the replication of the original effect. Thus, the EU Replication BF is ambiguous about the replication of the omnibus ANOVA effect (i.e., BF = .94), nor does it provide strong support for the replication of the post hoc result.

Note that the BFs according to the method presented in Ly et al. (2018) are higher than those calculated by the `ReplicationBF` package by Harms (2018b), even though both are extensions of Verhagen and Wagenmakers (2014). Harms (2018a) and Ly et al. (2018) discuss several differences between both approaches: (1)

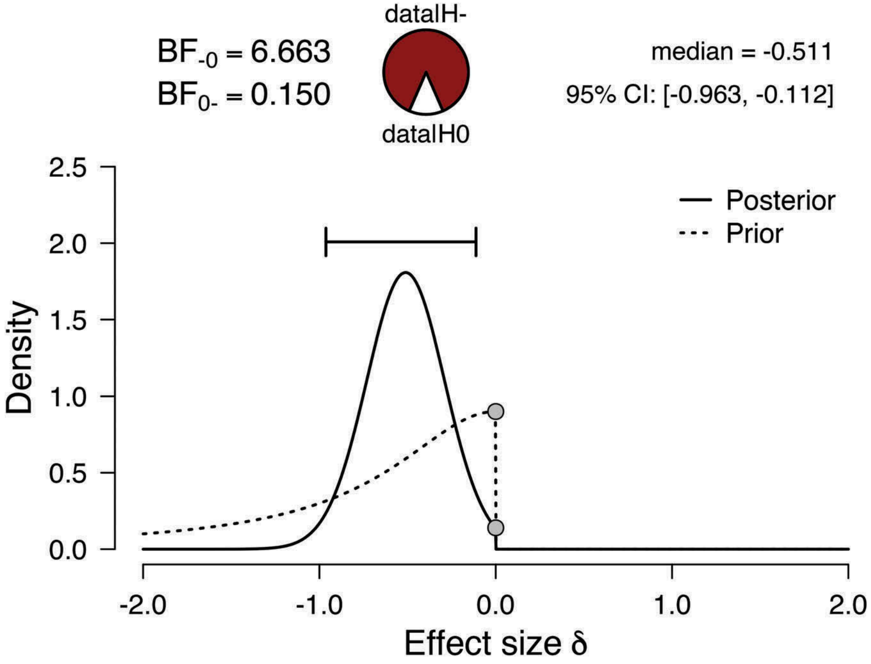

FIGURE 12.2 BF with default prior settings in the combined data for the one-sided *t*-test. The ratio of the two distributions at 0 on the x-axis is the BF

both methods use different priors (i.e., uniform in ReplicationBF R package, Cauchy in JASP and the BayesFactor R package), (2) the EU Replication BF assumes exchangeability to compute the BF for the combined data, and (3) for ANOVA models the BF computed in the ReplicationBF package is based on the sample size and test statistics, whereas JASP and the BayesFactor package use a more elaborate model that involves the full dataset(s). JASP currently also has a Summary Statistics module for *t*-tests, regression analyses, and analyses of frequencies. Whenever possible, we recommend applying both methods to obtain a more robust evaluation of replication.

Question 2: Is the new effect size different from the original effect size?

To test whether the new effect size is different from the effect size in the original study, we would preferably compute a confidence interval for the difference in effect sizes. The literature does not provide such an interval for η^2 or ω^2. However, with an iterative procedure based on descriptive statistics we can obtain separate confidence intervals for ω^2 in the original and new study (Steiger, 2004). Let us denote the original study with subscript *original*, and the new study with subscript *new*. For the original study $\omega^2_{original} = .09$, 95% CI [.00, .30] (see Supplementary Materials for all calculations). For the new study $\omega^2_{new} = .02$, 95% CI [.00, .22]. With these confidence intervals, we can calculate a confidence interval for the difference between both effect sizes, $\Delta\omega^2$, by applying the modified asymmetric method introduced by Zou (2007) for correlations and squared correlations. This method takes into account that some effect sizes have asymmetric distributions or cannot take on negative values (such as ω^2). $\Delta\omega^2 = .07$, 95% CI [-.15, .29]. Since zero is in the confidence interval of the difference between the effect sizes, we do not reject the hypothesis that the effect sizes are equal, and thus we retain the hypothesis that the new effect replicates the original one.

For the post hoc difference between the one-sided and two-sided implemental conditions we can compute the 95% confidence interval for standardized mean differences (i.e., Cohen's $d_{original} = .93$ and Cohen's $d_{new} = .36$) as given in Bonett (2009) and included in the Supplementary Materials. The difference between Cohen's *d* for both studies is .57, 95% CI [-0.96, 2.10]. Since zero lies in the confidence interval, we do not reject replication of the original effect size.

Alternatively, Patil, Peng, and Leek (2016) describe how non-replication of an effect size can be tested with a prediction interval. A 95% prediction interval aims to include the (effect size) estimate in the next study for 95% of the replications. Patil and colleagues (see Supplementary Materials) apply this method on *r* as calculated in the RPP. Following their methods, we find that the prediction interval for $r_{original} = .26$ ranges from -.12 to .57. The estimate for the new study, $r_{new} = .16$, lies within the interval of estimates that are expected given replication (i.e., -0.12 to 0.57). Hence, we do not reject replication of the original effect size. Note that Patil et al. apply their method on *r*, which is considered problematic when r is based on more than two groups (see, for example, Appendix 3 at osf.io/z7aux). The post hoc t-test value of $r_{original}$ is .42 with a prediction interval ranging from -0.03 to 0.73. For the new study, $r_{new} = .18$. Again, the correlation estimate for the new study lies within the prediction interval, and we do not reject the hypothesis that the original effect has been replicated.

The confidence intervals for the difference between effect sizes and the prediction intervals in this example can be considered to be quite wide. If the study results are uncertain (i.e., based on small samples), the associated confidence and prediction intervals will less often reject replication of the original effect size. However, especially with small studies, a failure to reject replication does not necessarily imply replication, but rather a lack of power, which suggests that the above methods may be inadequate for small samples.

Question 3: Are the new findings different from the original findings?

In contrast to the first two replication research questions which concerned effect sizes, the current question concerns study findings in general. The prior predictive p-value can be used to answer this question (Box, 1980; Zondervan-Zwijnenburg, Van de Schoot, & Hoijtink, 2019). The calculation of the prior predictive p-value starts with the simulation of datasets from the predictive distribution (with the sample size used in the new study) that are to be expected, given the original results. Subsequently, the new observed data from the replication attempt are compared to the predicted data with respect to relevant findings as summarized in H_{RF}. This hypothesis includes the relevant findings of the original study in an informative hypothesis (Hoijtink, 2012) and can include the ordering of parameters (e.g., $\mu_1 > \mu_2$), the sign of parameters (e.g., $\mu_1 > 0$, $\mu_2 < 0$), or the exact value of parameters (e.g., $\mu_1 = 3$, $\mu_2 = -2$). Any combination of constraints is possible. The deviation from the hypothesis for each of the predicted datasets and for the new dataset is expressed in the statistic that we call $\bar{F}$. Lower $\bar{F}$ values indicate a better fit with the relevant features specified in H_{RF}. With $\alpha = .05$, replication of the study findings is rejected if the misfit with H_{RF} in the new study is equal to or higher than in the extreme 5% of the predicted data. All computations can be conducted in an online interactive application presented at osf.io /6h8x3 or with the `ANOVAreplication` R package (Zondervan-Zwijnenburg, 2018). The online application (utrecht-university.shinyapps.io/anovareplication) and R package can take either raw data or summary statistics and sample sizes as input.

The results and conclusion of Henderson et al. (2008) lead to the following: $H_{RF} : \mu_{\text{One-sided implemental}} < (\mu_{\text{Two-sided implemental}}, \mu_{\text{Neutral}})$; Cohen's $d_{\text{One-sided implemental, Two-sided implemental}} > .8$.

If we run the test, we find that the prior predictive p-value = .130. Hence, we do not reject replication of the original study findings. Figure 12.3 shows the statistic $\bar{F}$ for each of the predicted datasets and the replication by Lane and

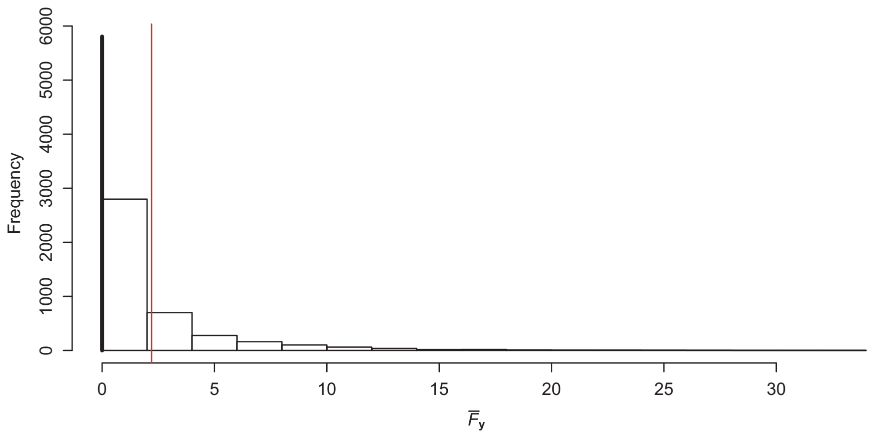

FIGURE 12.3 Prior predictive *p*-value. The histogram concerns $\bar{F}$ scores for each of the 10,000 predicted datasets with respect to the replication hypothesis. The thick black line represents the 5,805 predicted datasets that had an $\bar{F}$-score of exactly 0 and were perfectly in line with the replication hypothesis. The red line indicates the $\bar{F}$-score of 2.20 for the new study. The $\bar{F}$ score for the new data is positioned in the extreme 13.0% of the predicted data (prior predictive p = .130)

Gazerian (2016). Note that we do not have to run a post hoc analysis with this method, because the conclusion for the post hoc contrast was incorporated in H_{RF} with "Cohen's $d_{\text{One-sided implemental, Two-sided implemental}} >$ .8". For the prior predictive *p*-value, an original study with large standard errors (e.g., due to a small sample) leads to a wide variety of predicted datasets, thus making it hard to reject replication of the original study conclusions. With the `ANOVAreplication` `R` package we can calculate the power to reject replication when all means would be equal in the new study. Here, the statistical power was only .66, which means that the probability that we do not reject replication incorrectly (i.e., a Type II error) is 1-.66=.34. When we calculate the required sample size to obtain sufficient power, we find that a sample size of 41 per group would be required to reject replication of H_{RF} in a sample with equal group means.

Question 4: What is the effect size in the population?

At the end of the day, most researchers are concerned with the effect in the population. To determine the population effect based on an original and new study, numerous meta-analytic procedures have been proposed. For close replications, the fixed-effect meta-analysis can be used, which assumes that there is one underlying population from which both studies are random samples. Consequently, there is only one underlying true effect size. However, the standard fixed-effect meta-analysis does not take publication bias into account. As a result, standard fixed-effect meta-analyses overestimate effect sizes.

The frequentist hybrid meta-analysis (Van Aert & Van Assen, 2017b), and the Bayesian snapshot hybrid method (Van Aert & Van Assen, 2017a) are two meta-analytic methods developed for situations with a single replication effort that take into account the significance of the original study (which could be caused by publication bias). Both methods are part of the `puniform` `R` package (Van Aert, 2018) and available as online interactive applications. The required input are descriptive statistics and effect sizes. The frequentist hybrid meta-analysis results in a corrected meta-analytic effect size and its associated confidence interval and *p*-value. The output also includes the results of a standard fixed-effect meta-analysis for comparison. The Bayesian snapshot hybrid method quantifies the relative support, given the original and replication study, for four effect size categories: zero, small, medium, and large. Currently, both methods can be used for correlations and *t*-tests. However, the correlation for the original ANOVA as computed by the RPP cannot be used for the meta-analytic methods, because its standard error cannot be computed for more than two groups.

For the post hoc *t*-test results of Henderson et al. (2008) and Lane and Gazerian (2016), the bias-corrected Hedges' *g* is .37, 95% CI [-0.48, 0.94], $p = .232$. Thus, we cannot reject the hypothesis that the effect in the population is zero. The standard (uncorrected) fixed-effect meta-analytic estimate was .60, 95% CI [0.12,1.07], $p = .014$. Whereas the fixed-effect meta-analytic effect sizes was significant at $\alpha = .05$, the hybrid meta-analysis effect size is lower and has a wider 95% confidence interval. The snapshot hybrid method with equal prior probabilities for the four effect size categories indicated that a small effect size received the highest support (37.8%), followed by no effect size (30.2%), a medium effect size (25.5%), and a large effect size (6.6%).

Besides meta-analyses that take significance of the original study into account, we can also calculate the BF for an effect versus no effect, based on the scaled combined data using `JASP`. The BF in favor of an ANOVA effect is 1.50. The BF in favor of a post hoc *t*-test effect is 6.66. Hence, the evidence in the combined data is positive with respect to the existence of an effect. Note that this combined analysis does not correct for publication bias and assumes exchangeability. Alternatively, Etz and Vandekerckhove (2016) developed a BF for *t*-tests, univariate *F*-tests (i.e., for up to two groups), and univariate regression analyses that takes into account publication bias, but unfortunately this BF has only been developed for the `MATLAB` software package, which is not commonly used in the social sciences and will not be described further.

Discussion

In this chapter, we presented replication research questions and associated statistical techniques. In the example we used, the replication BFs pointed mostly towards a null effect instead of a replication of the original effect; the confidence intervals around the difference between effect sizes indicated that the difference between the original and new study may be zero, but that they had low power;

the prior predictive *p*-value demonstrated non-replication of the original study conclusions; and meta-analyses indicated that the population effect is small, anecdotal, or not significantly different from zero.

We also discussed how the different methods perform with small samples. BFs and the Bayesian snapshot meta-analysis have the advantage over null-hypothesis significance testing (NHST) methods (e.g., confidence intervals and the prior predictive *p*-value) that they cannot be underpowered. The evidence by the BF may not be overwhelming, but at least it indicates the relative plausibility of one hypothesis over the other after observing the data. With additional data and Bayesian updating methods (see also Chapter 9), the evidence can become more convincing. NHST methods, on the other hand, often result in non-significant findings with small samples, and it remains unclear whether the (non)replication effect was absent, or whether the analysis was underpowered.

An advantage of the prior predictive *p*-value is that it allows the user to test the replication of the original study findings summarized in H_{RF}. This hypothesis can include multiple parameters, and it can convey information on their size and ordering. In the ANOVA setting, the effect size (e.g., η^2) does not provide information about the direction of the effect. Hence, it is useful to evaluate relevant features that can cover the ordering of group means.

The preferred method to test replication depends on the replication research question at hand. Furthermore, given a replication research question, it can be insightful to apply multiple methods to test replication (Harms, 2018a). Testing replication yields more meaningful results with larger sample sizes, and this holds for all methods described in this chapter. Testing replication of small sample research is challenging, but since small samples are more susceptible to researcher degrees of freedom, it is of utmost importance to critically evaluate small sample results with replication studies.

Author note

The first author is funded through the Consortium Individual Development by a Gravitation program grant of the Dutch Ministry of Education, Culture and Science, and the Netherlands Organization for Scientific Research (grant number 024.001.003).

Note

[1] We report BF up to two decimal places, but use all available information for calculations.

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.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, *14*(3), 225–238. doi:10.1037/a0016619.

Box, G. E. P. (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.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology*, *65*(3), 145–153. doi:10.1037/h0045186.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One*, *11*(2), e0149794. doi:10.1371/journal.pone.0149794.

Harms, C. (2018a). A Bayes factor for replications of ANOVA results. *American Statistician*. doi:10.1080/00031305.2018.1518787.

Harms, C. (2018b). ReplicationBF: Calculating replication Bayes factors for different scenarios. Retrieved from https://github.com/neurotroph/ReplicationBF.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. doi:10.3102/10769986006002107.

Henderson, M. D., De Liver, Y., & Gollwitzer, P. M. (2008). The effects of an implemental mind-set on attitude strength. *Journal of Personality and Social Psychology*, *94*(3), 396–411. doi:10.1037/0022-3514.94.3.396.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: Taylor & Francis.

JASP Team. (2018). JASP Version 0.9.1.

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

Lane, K. A., & Gazerian, D. (2016). Replication of Henderson, De Liver, & Gollwitzer (2008, JPSP, Expt. 5). osf.io/79dey.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. doi:10.1177/0956797615616374.

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.

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs [R package version 0.9.12-4.2]. Retrieved from https://CRAN.R-project.org/package=BayesFactor.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129–147. doi:10.2333/bhmk.40.129.

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

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.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi:10.1037/0033-2909.86.3.638.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632.

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

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182. doi:10.1037/1082-989X.9.2.164.

Van Aert, R. C. (2018). Puniform: Meta-analysis methods correcting for publication bias [R package version 0.1.0]. Retrieved from https://CRAN.R-project.org/package=puniform.

Van Aert, R. C., & Van Assen, M. A. (2017a). Bayesian evaluation of effect size after replicating an original study. *PLoS One*, *12*(4), e0175302. doi:10.1371/journal.pone.0175302.

Van Aert, R. C., & Van Assen, M. A. (2017b). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, *50*(4), 1515–1539. doi:10.3758/s13428-017-0967-6.

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.

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*, 192–196. doi:10.1007/s11606-010-1513-8.

Zondervan-Zwijnenburg, M. A. J. (2018). ANOVAreplication: Test ANOVA replications by means of the prior predictive p-value. Retrieved from https://CRAN.R-project.org/package=ANOVAreplication.

Zondervan-Zwijnenburg, M. A. J., Van de Schoot, R., & Hoijtink, H. (2019). Testing ANOVA replications by means of the prior predictive p-value. doi:10.31234/osf.io/6myqh.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. doi:10.1037/1082-989X.12.4.399.