Assessing Evidence for Replication: A Likelihood-Based Approach

Peter Dixon

University of Alberta

and

Scott Glover

Royal Holloway University of London

Abstract

How to evaluate replications is a fundamental issue in experimental methodology. We develop a likelihood-based approach to assessing evidence for replication. In this approach, the design of the original study is used to derive an estimate of a theoretically interesting effect size. A likelihood ratio is then calculated to contrast the match of two models to the data from the replication attempt: 1) A model based on the derived theoretically interesting effect size; and 2) a null model. This approach provides new insights not available with existing methods of assessing replication. When applied to data from the Replication Project (Open Science Collaboration, 2015), the procedure indicates that a large portion of the replications failed to find evidence for a theoretically interesting effect.

Assessing Evidence for Replication: A Likelihood-Based Approach

There has been a great deal of concern expressed recently regarding the "replication crisis" in psychology (e.g., Lindsay, 2015; Pashler & Harris, 2012; Shrout & Rodgers, 2018), wherein a potentially large number of published results may be difficult to replicate (Camerer, Dreber, Holzmeister, et al, 2018; Klein, Vianello, Hassleman et al., 2018; Open Science Collaboration, 2015). Low replication rates have been ascribed to a number of factors, including data analysis strategies that inflate the Type I error rate (e.g., Bishop, 2019; Simmons, Nelson, & Simonsohn, 2011), publication practices that favor reporting significant results (e.g., de Bruin, Treccani, & Della Sala, 2015; Francis, 2012), and inherent problems with significance testing (e.g., Masicampo & Lalande, 2012; Wassersman & Lazar, 2016). Any or all of these issues may indeed contribute to a failure to replicate, but an equally important question revolves around what counts as evidence for or against replication. In fact, it seems crucial to have a solid statistical foundation for deciding whether a replication has been successful or not before addressing issues related to improving replicability itself.

In the present paper, we argue that there are different senses in which a result may or may not replicate. As an illustration, we contrast two recently offered approaches to replication, a Bayes factor test proposed by Verhagen and Wagenmakers (2014) and the "small-telescopes" approach of Simonsohn (2015). Following this, we describe a likelihood-based approach to replication based on the evidence for what might be the theoretically interesting effect size implicit in the original study. As an illustration of the technique, we apply this new approach to data from the Reproducibility Project (Open Science Collaboration, 2015). We conclude that our method provides important new insights into the assessment of replication that are not available with other approaches.

The Aims of Replication

A core problem in science is deciding whether or not an observed result provides evidence for a theoretically interesting effect. As many have noted, a theoretically interesting effect is not the same as a statistically significant effect (e.g., Thompson, 1993). For example, an effect of any magnitude can be statistically significant given sufficient power, whereas an effect generally must be of a certain magnitude to be considered theoretically interesting. Further, there is good reason to believe that in many paradigms, there is always going to be some minimal difference between conditions for reasons that have little to do with the question of interest (cf. Bakan, 1966; see also Meehl, 1990). Thus, effects must be of some minimal value to provide a meaningful insight into the research question. We assume that published papers should generally report reasonable evidence for theoretically interesting effects given that such evidence is a central criterion on which publication depends.

From the perspective of the field and for the advancement of scientific knowledge, it is unimportant whether a replication produces precisely the same result as the original study. Rather, what matters is whether the replication evidence supports the same *interpretation* as the original, namely that evidence exists for a theoretically interesting effect. Thus, an important aspect of a replication attempt is an answer to the question: Does the evidence from the replication support the existence of a theoretically interesting effect or not? That said, the magnitude that an effect must have to be theoretically interesting may be difficult to determine. Although researchers may have an intuitive understanding of this magnitude, it is rarely discussed in research reports. Further, information about the variability of an effect in the population (and hence the standardized effect size) might be lacking for novel findings or paradigms. The technique we develop below provides a way to estimate the size of the theoretically interesting effect that might have been anticipated by the original researchers by examining the design of their study.

A second aspect of assessing evidence for replication is that one's concerns are often symmetrical: We generally wish to know both when the evidence is in favor of replication and when it is against. If one can gauge the magnitude of a theoretically interesting effect, the question can be posed in this symmetrical fashion. That is, we can ask: Does the replication evidence better support the existence of a theoretically interesting effect, or does it better support a null effect? One benefit of such a symmetrical question is that it is straightforward to use likelihood ratios to describe the statistical evidence.

Techniques for Describing Replication

Both Verhagen and Wagenmakers (2014) and Simonsohn (2015) identified a number of problems with common approaches to evaluating replication. For example, comparing patterns of significance is problematic because one result may be significant and another nonsignificant even though the two effect sizes are comparable. Similarly, testing for a significant difference between the size of the original effect and that found in a replication attempt can be biased because null results are likely if the original effect is imprecise. These authors proposed solutions to these problems, and the present approach builds on these solutions.

Bayes-Factor Replication Test. Verhagen and Wagenmakers (2014) proposed a Bayesian approach to replication in which two possible interpretations of the replication's effect size are compared: The first is that the replication results are consistent with the posterior distribution derived from the original study; the second is a null model in which the effect is assumed to be zero. This technique has the advantage of framing the replication question in a symmetrical fashion: Are the replication results more consistent with the original study or with a null effect? Thus, it can provide evidence both for and against replication. As well, the evidence from the technique is expressed as a Bayes factor, which allows one to gauge the strength of the evidence in the context of the studies in question. However, it can still be difficult to find evidence against replication using this method when the original finding is imprecise. This is because the posterior distribution derived from an imprecise original study will be relatively diffuse, with some likelihood assigned to even small values of the effect.

"Small Telescopes." Simonsohn (2015) described an interesting alternative solution to the question of replication. Rather than assessing evidence for or against a previously obtained result, he argued that one should consider the magnitude of the effect one could reasonably be expected to find given the design of the original study. The first step here would be defining a "small effect" as an effect that could be found 33% of the time given the sample size used in the original study (described as "d_{33}"). Following this, an analysis would be conducted to see if the effect obtained in the replication attempt was significantly smaller than d_{33}. If the null hypothesis of no difference were rejected, one could conclude that the original result failed to replicate in the sense that the effect was smaller than what the original experiment could reasonably have been expected to find. In other words, the conclusion would be that the original experiment was "too small a telescope" to see the effect that was obtained.

This approach makes it easier to find evidence for a failure to replicate because rejecting a small effect specified *a priori* can be easier than finding evidence against an originally imprecise finding. However, the small telescopes approach requires that a relatively large sample is needed to find evidence against replication in many cases; for example, Simonsohn (2015) recommended samples 2.5 times as large as the original study. Further, the procedure is essentially a means to provide evidence that would discredit the original study, and a failure to reject the null hypothesis in this case does not provide clear evidence *for* replication. As such, this method doesn't fulfill the goal of assessing the evidence in a symmetrical fashion. In order to provide evidence for replication, one might conceivably supplement the small-telescopes test with a conventional test of significance against the null hypothesis of 0; however, such an approach requires special handling because the two significance tests are not independent.

In sum, both the Verhagen and Wagenmakers (2014) and Simonoshn (2015) methods provide improvements over naive methods of assessing replication, and both provide information that is valuable for interpreting the results of a replication attempt. Here, we build on these ideas to develop a likelihood-based approach that poses the question somewhat differently: Does the replication attempt provide evidence for a theoretically interesting effect, or is the evidence more consistent with a null effect?

A Likelihood-Based Approach

Our method for assessing evidence for replication combines elements of the Bayesian and the small-telescopes approaches to provide clear inferences concerning replication without the need for large samples. This approach uses the design of the original research to make a best guess as to how large a theoretically interesting effect would be. This depends on what we refer to as the "researcher-insight" assumption: that the original researchers had some insight regarding how large an effect would be theoretically interesting and that they designed a suitably powerful study based on that insight. Using this assumption and working backwards from the size of the original study, we can then make an informed estimate as to the magnitude of the effect for which the study might have been designed. Although this researcher-insight assumption may be debatable in many cases, we argue that it provides a useful starting point in assessing evidence for replication. Further, even if the assumption is incorrect, the results of the assessment can constrain further reasoning about the theoretically interesting effect size. Alternatively, in cases in which where the theoretical and empirical issues are relatively well understood, it may be possible to identify a suitable estimate of the theoretically interesting effect size with an *a priori* analysis, without depending on the researcher insight assumption. The present approach applies primarily to situations in which such analysis is not available.

In the present development, we build on the approach to assessing evidence described by Glover and Dixon (2004). There, we suggested using an “adjusted” likelihood ratio (λ_{adj}) to describe the evidence for one model of the data relative to another. The likelihood ratio is the likelihood of the data given one model of the results relative to the likelihood of the data given a competing model and can be written as:

$$\lambda = \frac{L_1}{L_0} \tag{1}$$

where L_0 and L_1 are the likelihoods given the two models. The adjusted likelihood ratio uses the Akaike (1973) Information Criterion to compensate for the fact that the evidence will nearly always favor the model with more parameters. Using a small sample approximation to the AIC yields the following expression for the adjusted likelihood ratio:

$$\lambda_{adj} = Q_c(n)\lambda \tag{2}$$

where

$$Q_c(n) = exp\left[k_2\left(\frac{n}{n-k_2-1}\right) - k_1\left(\frac{n}{n-k_1-1}\right)\right] \tag{3}$$

and k_1 and k_2 are the number of parameters in the two models. Such an adjusted likelihood ratio is tantamount to selecting models based on AIC values. Burnham and Anderson (2002) refer to such adjusted likelihood ratios as “evidence ratios.” The AIC-adjusted likelihood ratio is closely and inversely related to *p* values in some simple hypothesis-testing contexts but differs in that it provides an index of the relative strength of the evidence for two competing models rather than supporting a dichotomous accept/reject decision.

The first step in our procedure is to use the sample size of the original study to calculate the theoretically interesting effect size that might have been anticipated by the researchers. This is the minimum size of an effect for which the study could have been expected to produce good evidence. We assume that “good evidence” corresponds to an adjusted likelihood ratio of 8:1. (Although arbitrary to some extent, this criterion is somewhere in between weak evidence and very strong evidence. Somewhat different choices are possible, but this would not change the substance of our approach.) In a significance testing framework, this would correspond to a power of about .7 (as shown in the supplementary materials). We refer to this as the anticipated evidence for the study, $\lambda_{ae} = 8$, and the corresponding effect size as the anticipated theoretically interesting effect size, d_{tie}. With normal data, the likelihood of the data is a simple function of variance that is unexplained by a model. This allows one to calculate d_{tie} from the sample size with some algebraic manipulation:

$$d_{tie} = 2\left[\frac{n-2}{n}\left(\left[\frac{8}{Q_c(n)}\right]^{2/n} - 1\right)\right]^{1/2} \tag{4}$$

where *n* is the original sample size. (These calculations are derived in the supplementary materials.) Of course, this calculation would not be necessary if d_{tie} could be identified via an analysis of the research domain. Indeed, if such an analysis leads to a value that is substantially less than the results of Equation 4, one might conclude that the original study was underpowered.

Having determined d_{tie}, we then consider how well the effect obtained in a replication attempt, d_{obt}, is explained by two models: a null model assuming no effect and a replication model that assumes the effect is d_{tie}. A likelihood ratio is used to describe how likely the obtained effect is given the replication model relative to how likely it is given the null model. As shown in the supplementary materials, this ratio is:

$$\lambda_{rep} = \frac{L_{rep}}{L_0} = \left[\frac{\frac{n}{n-2}\left(\frac{d_{obt}}{2}\right)^2 + 1}{\frac{n}{n-2}\left(\frac{d_{tie} - d_{obt}}{2}\right)^2 + 1} \right]^{\frac{n}{2}} \qquad (5)$$

where n is the sample size in the replication attempt. The magnitude of the likelihood ratio describes the strength of the evidence in favor of one or the other model. Very large ratios in favor of the anticipated theoretically interesting effect would be considered strong evidence *for* replication. Symmetrically, very large ratios in favor of the null model would be strong evidence *against* replication. Smaller ratios in either direction would be weaker evidence, and ratios near 1:1 would be inconclusive.

Our approach is illustrated in Figure 1. In this example, we assume that in the original study, there were 40 participants in a between-participants design with two groups of 20. From Equation 4, the anticipated theoretically interesting effect size is $d_{tie} = 0.82$. We assume that a replication attempt used a somewhat larger sample with 60 participants (again, with two groups of 30). The solid curve indicates the likelihood corresponding to different possible effect sizes in such a replication attempt if the true effect size was equal to $d_{tie} = 0.82$. The dashed curve indicates the likelihood corresponding to possible effect sizes if the true effect size were 0. The solid and dashed gray vertical lines depict the ratio of these two likelihoods under three different scenarios. On the far right, it is assumed that the replication attempt found an effect size of d_{obt} = 0.70; this is much more likely given the anticipated-evidence model than the null model and produces a likelihood ratio of λ_{adj} = 32.07, or compelling evidence in favor of replication. In the center, the replication attempt produced an effect size of d_{obt} = 0.50. This produces a likelihood ratio of only λ_{adj} = 2.99. Although it favors replication, this is fairly weak evidence. Finally, on the left, an effect size of d_{obt} = 0.25 was obtained in the replication attempt. This is more likely under the null model than under the anticipated-evidence model and leads to a likelihood ratio of λ_{adj} = 0.14 or, inversely, λ_{adj} = 6.89 in favor of failure to replicate. (These example calculations are detailed in the supplementary materials.)

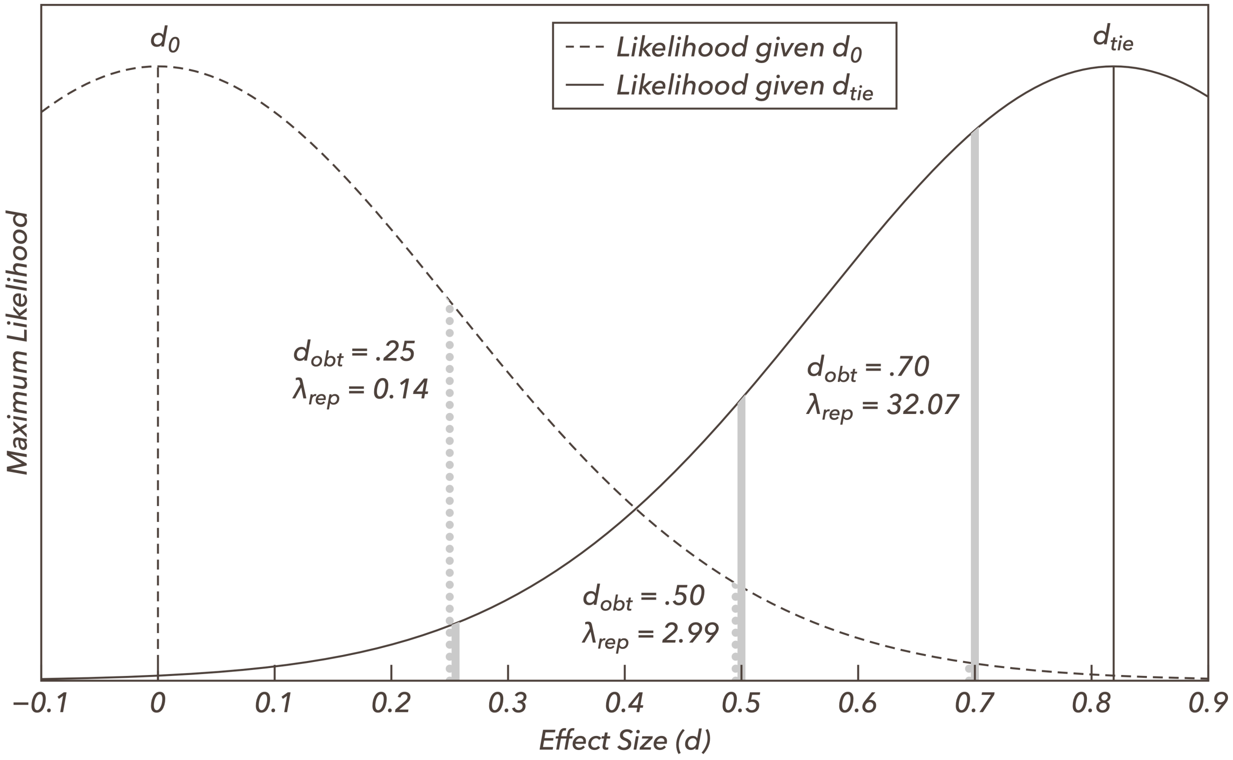

Figure 1. Likelihood ratios under the null and anticipated-effect models for three different obtained effects. The solid curve represents the expected distribution of effect sizes under the anticipated effect model assuming a theoretically interesting effect, while the dotted curve represents the same under the null model.

This technique provides an effective way to describe the evidence for or against replicating a theoretically interesting effect size. However, “failure to replicate” has a specialized interpretation in this context: It means that the obtained estimate of the effect is smaller than a theoretically interesting effect (based on the researcher-insight assumption) and is better described as being equal to zero. Because of this specialized interpretation, one possible conclusion following from evidence for a “failure to replicate” is that the estimate of the theoretically interesting effect size is exaggerated. Indeed, it is possible with this outcome that the design of the original study was not sufficiently powerful to detect a small (but potentially theoretically interesting) effect. Our view is that if there is substantial evidence against replication using this procedure (and even when there is weak evidence for replication), it should prompt a careful analysis of the research paradigm and theoretical context in order to arrive at a deeper understanding of how large an interesting effect would be. In many cases, a natural conclusion may be that more powerful studies would be needed to detect a smaller effect. Thus, our researcher-insight assumption, although not always justified, can in many cases lead to the identification of contexts in which further analysis of a theoretically interesting effect size is required.

Comparison to Other Approaches

The present approach has similarities with both the Simonsohn (2015) small-telescopes approach and the Verhagen and Wagenmakers (2014) Bayes-factor replication test outlined in the introduction. Regarding Simonsohn (2015), our likelihood-based approach is similar in that the comparison is between results of the replication attempt and an index derived from the design of the original study (rather than the actual results of that study). However, the present approach differs from that of Simonsohn in four important respects: First, the approach can also provide positive evidence for successful replication (which does not directly follow from the small-telescopes test). Second, our approach provides a continuous index of the strength of the evidence for or against replication, rather than the dichotomous decision that emerges from significance testing. (Using the obtained *p* value as an index of strength of evidence would be inconsistent with the tenets of significance testing as commonly described.) Third, because the comparison is between two *a priori* point alternatives, a more sensitive index of failing to replicate is possible, typically with about half the sample size for comparable levels of power (see supplementary materials). Fourth, there is an important conceptual difference in what is entailed by "failure to replicate." In the small-telescopes test, this means rejecting the null hypothesis that the effect was large enough to be readily detectable in the original study. In the likelihood-based approach, this means that the null model provides a better account of the replication data than a model based on the theoretically interesting effect size determined by the size of the original study. This distinction can easily lead to different interpretations. For example, the likelihood-based approach might imply that a small replication effect size provides evidence for a failure to replicate if the sample size of the original study was small (implying a large anticipated effect). In contrast, given the same data, the small-telescopes method could easily fail to reject the null hypothesis, resulting in an inconclusive interpretation.

Regarding the Verhagen and Wagenmakers (2014) approach, the likelihood-based analysis is similar in that both pose the question symmetrically in terms of which model is supported by the evidence (the "replication" model versus the null). Thus, both methods allow for evidence either for or against replication. However, a critical difference is that we compare the null model to a model based on a theoretically interesting effect size rather than the posterior distribution estimated from the original results. This difference can lead to different insights regarding a potential failure to replicate. For example, as shown in Figure 1, according to our method, when a small observed effect size of d_{obt} = 0.25 is compared to the anticipated effect size of d_{tie} = 0.82, there is clear evidence of 6.95:1 against replication. However, in the Verhagen and Wagenmakers approach, such data produce equivocal results, with a Bayes factor of 0.54 in favor of replication, or 1.84 in favor of failure to replicate. More interestingly, if that same small effect size were observed in a very large replication attempt with five hundred observations, the Bayes factor would actually favor replication, whereas the likelihood-based approach would provide very strong evidence *against* replication.

We believe that this difference in interpretation arises because of Verhagen and Wagenmakers are comparing the null model to the posterior distribution of the effect based on the original results. Because the posterior distribution of the effect size is, to some extent, diffuse, it will have at least some density at even small values of the effect size. Thus, a very precise, but small, estimate of the effect size can be more consistent with the posterior distribution than with zero. In contrast, in the likelihood-based approach we are comparing *two a priori* point values, zero and the effect size based on the original study's design. Thus, a small effect is likely to be more consistent with zero than with a large anticipated effect, regardless of sample size. Under such circumstances, it would be reasonable to conclude that the replication attempt produced an effect that is much smaller than what the researchers might have originally expected. Thus, a researcher using our method might reasonably be led to ponder whether the effect is real but smaller than the effect suggested by the design of the original study. The researcher would, of course, then have to evaluate whether or not this new estimate of effect size was of sufficient magnitude to be considered theoretically interesting.

Application to the Reproducibility Project

As an illustration of the likelihood-based approach, we applied it to results from the Reproducibility Project (2015). The project reported the results of 100 attempts to replicate quasi-randomly selected research results from across a broad range of psychology journals. Their results are important because they provide an independent assessment of the extent to which results in psychology are replicable. For simplicity, we considered studies for which the relevant test statistic was either *t* or *F* (although the current approach could be extended to other analyses). We also did not use studies for which the original or the replication attempt had a sample size greater than 1,000 because these would be atypical of replication attempts in experimental psychology. We also omitted one additional study as atypical because the effect degrees of freedom was 18. This resulted in a total of 83 pairs of studies. For each pair, we calculated the anticipated effect size for the original design and the replication effect size from the reported test statistic. We then calculated the evidence for or against replication. For the purposes of this application, we made two adjustments to the approach developed so far. First, because some portion of the studies involved effects with more than a single degree of freedom, we calculated effect size in terms of f^2 rather than d. Second, we describe the evidence as the difference in AIC values for the null and replication models, effectively putting the likelihood ratio on a log scale. (Although we regard likelihood ratios as more intuitive, the AIC difference is more suitable for a graphical presentation of evidence for and against replication.) In this case, large positive values of the difference in AIC values would indicate evidence for replication and large negative values against replication.

The results are shown in Figure 2; details are provided in the supplementary materials. A value of 3 for ΔAIC might be considered as clear evidence under many circumstances, and this criterion is shown as dotted lines in the figure. (For example, in some prototypical hypothesis testing situations, an obtained *p* value of .05 corresponds to a ΔAIC of 2.2.) It is notable that using our likelihood-based replication assessment, there was evidence against replication in a large portion of the results. In this sense, the present approach does not change the broad conclusions from the Reproducibility Project, although we believe that these calculations provide additional insight into the problem.

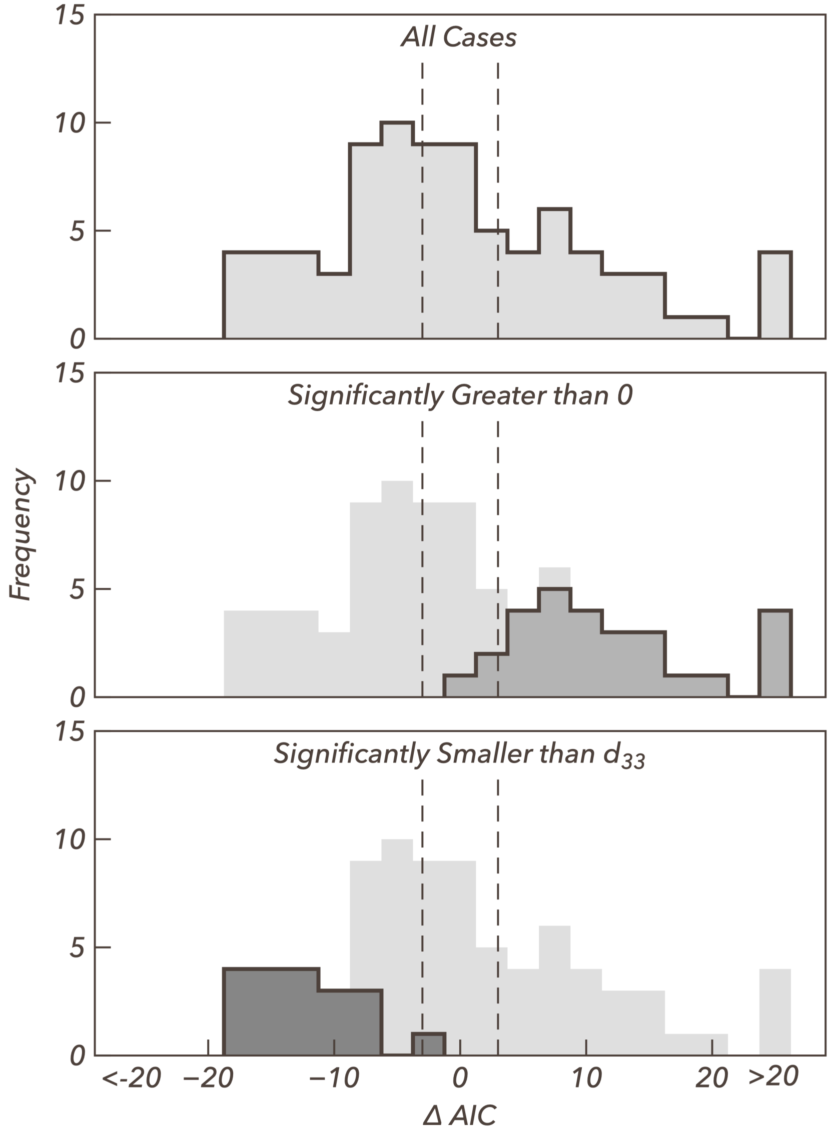

Figure 2. Results of applying the likelihood-based approach to studies in the Reproducibility Project (Open Science Collaboration, 2015). In the top panel, gray areas indicate the frequency of ΔAIC (difference in AIC values; see Equation 2.10 in the supplementary materials), and dotted vertical lines indicate the criteria of ±3 ΔAIC. The dark gray in the center panel depict the subset of studies for which the effect of interest was significantly greater than 0. The dark gray areas in the bottom panel depict the subset for which the effect was significantly smaller than d_{33} (Simonsohn, 2015).

As a comparison to other indices of replication, two other measures are shown in Figure 2. The first was whether or not the result in the replication attempt was statistically significant. Significant replication attempts are indicated by the dark gray area in the center panel. As can be seen, most of the cases in which the likelihood-based assessment yielded clear evidence for replication were also statistically significant. However, there were a few instances in which a significant effect was found but there was only weak evidence for replication. This can occur when the replication attempt has substantially higher power than the original study. Under such circumstances, the replication attempt may find a small, significant effect that is substantially smaller than the anticipated effect size estimated from the design of the original study. Note as well that a failure to find a statistically significant effect using standard significance testing did not always correspond to evidence for a failure to replicate using our approach.

In the second comparison, the results of the likelihood-based approach were compared to those results that represented a failure to replicate using the small-telescopes approach of Simonsohn (2015). These are depicted as the dark gray area in the bottom panel. As can be seen, those results in which the small-telescopes approach indicated a failure to replicate were also failures to replicate using the likelihood-based approach. In other words, when the obtained effect was smaller than d_{33} (using a significance test), it was also smaller than d_{tie} (and was better fit by a null model). This is perhaps not surprising given that both criteria are derived from the sample size of the original study. However, there were also number of cases for which the likelihood-based approach suggested a failure to replicate when the small-telescopes approach did not. This highlights the different interpretation of "failure to replicate" in the two approaches. For the small-telescopes approach, "failure to replicate" means rejecting the hypothesis that the effect is larger than a small effect (estimated from the original study); for the likelihood-based approach, “failure to replicate” means that the effect is better fit by null model than a model assuming a theoretically interesting effect size (estimated from the original study).

We also compared the results from the likelihood-based approach to the replication Bayes factor approach (Figure 3). Here, we plotted the difference in AIC values against the log of the Bayes factor. These two measures need not be precisely related since the Bayesian approach depends on the actual results of the original study while the likelihood-based approach depends on the anticipated effect size. Nevertheless, there is a clear relationship. The largest difference between the two patterns of results is that the Bayes factor is generally not as strong as the difference in AIC values, particularly for evidence against replication. This highlights our observation that for similar sample sizes, it is more difficult to find evidence against replication using the Bayes factor approach. We conjecture that this difference arises because the posterior distribution given the original result can be diffuse. This means, for example, that even small obtained effects have some likelihood given the posterior, and there is thus a limit in how small the Bayes factor can be. In contrast, the likelihood-based approach compares two point hypotheses (0 and d_{tie}), and it is quite possible for an obtained effect to be much more consistent with one than with the other. There is also a discrepant result for one study that had a slightly positive AIC difference but a large negative Bayes factor. The original study in this instance had a large sample size (and consequently a small value for d_{tie}), but also a very large effect size. The replication attempt found a small effect between d_{tie} and 0. Consequently, the likelihood-based approach indicated that the replication was inconclusive. However, the Bayesian approach demonstrated that the obtained result was smaller than the original obtained result. This highlights the difference in questions being asked: The likelihood-based approach is concerned with whether the replication attempt found evidence for a theoretically interesting effect; the

Bayesian approach is concerned with whether the effect is similar to that obtained in the original study.

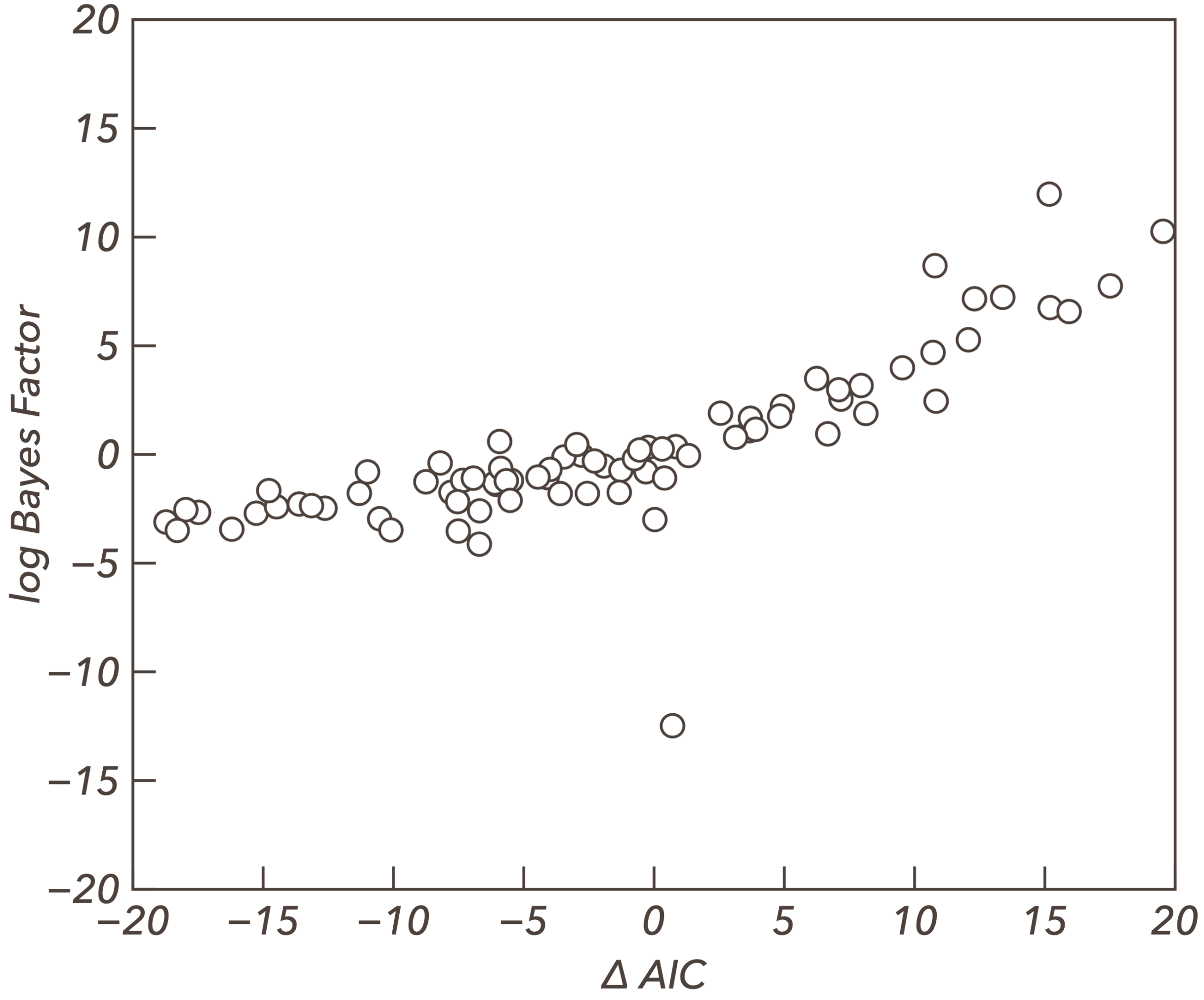

Figure 3. Relationship between log Bayes factor and AIC difference when applied to data from the Reproducibility Project (Open Science Collaboration, 2015). For clarity, four studies with an AIC difference greater than 20 are not shown.

Limitations

While we argue that the present approach can provide insights into the question of replication, there are a variety of constraints on our conclusions. As with all statistical models, the validity of the conclusions depends on the accuracy of the underlying assumptions. For example, our derivations are predicated on independent, normally distributed data. We are uncertain whether deviations from this assumption would lead to a bias towards the null or replication model in Equation 4. As noted above, the present development is limited to the simple comparison of two conditions; the extension to more complex statistical questions is possible but not treated here. Further, our analysis depends on the choice between two point estimates for the mean, with no variation in other aspects of the distributions. In some applications, such constraints may be unreasonable. Nevertheless, we feel that there is no substantial obstacle to applying the general approach to a broad range of other situations.

Concluding Comments

Our likelihood-based approach has several important differences from other methods of assessing replication. First, it poses the question of replication symmetrically, so that evidence can be found either for or against replication. This is in contrast to other methods which often can only answer one or the other side of the question. Second, it allows for a graded and intuitive description of the evidence. This avoids some of the problems with null-hypothesis significance testing that derive from the use of an arbitrary decision-making criterion. Finally, it uses the provisional, working assumption that the original researchers designed their study to be sufficient to detect a theoretically interesting effect, and it infers the expected effect size based on their design rather than on the obtained effect. Although this researcher-insight assumption may often be incorrect, the approach can nonetheless provide a starting point for evaluating the evidence for replication. Further, where that assumption appears to fail, it may encourage a more careful evaluation of what might be considered an effect size sufficient to be considered theoretically interesting within that specific paradigm.

Although we have developed and applied this technique in terms of likelihood ratios, the same concepts could be used regardless of one's approach to assessing evidence. In particular, using the design of the original experiment to assess the original researcher's expectations does not depend on any assumptions about how competing hypotheses should be compared. For example, the same approach could be used by starting with the significance-testing concept of power and then developing mutually exclusive point hypotheses. As another example, a Bayesian version of the procedure could be developed by using the Bayesian model comparison statistic BIC instead of AIC in all of the present developments. Although these alternative approaches would, in general, be numerically different when applied to specific cases, we believe that the conclusions would typically be similar.

Understanding why results cannot be replicated is a critical issue in psychology and other sciences, but assessing replicability is equally crucial, for without a robust evaluation of the problem one cannot formulate suitable solutions, especially in terms of assessing individual attempts to replicate. Our approach is to focus on the central question, "Does the data provide evidence for a theoretically interesting effect, or for a null effect?", and to frame this question in a symmetrical fashion. This approach in conjunction with the use of likelihood ratios allow a graded and principled means of assessing replication.

Open Practices Statement

Supplementary materials, including R code used for generating simulations and analyses, are available at https://osf.io/xfy53/download

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd international symposium on information theory* (pp. 267-281). Budapest: Akademia Kiado.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . Zuni, K. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science (New York, N.Y.)*, *351*(6277), 1037. doi:10.1126/science.aad916

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423-437.

Bayarri, M. J., & Mayoral, A. M. (2002). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, *103*(1), 225-243. doi:10.1016/S0378-3758(01)00223-3

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature, 568,* 435.

de Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science : A Journal of the American Psychological Society / APS*, *26*(1), 99-107. doi:10.1177/0956797614557866

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. New York: Springer.

Camerer, C. F., Dreber, A., Holzmeister, F., et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour, 2,* 637-644.

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

Dixon, P. (2003). The *p* value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*, 189-202.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin and Review, 19,* 151-156.

Gigerenzer. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587-606.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science (New York, N.Y.)*, *351*(6277), 1037. doi:10.1126/science.aad724

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791-806.

Klein, R. A., Vianello, M., Hasselman, F. et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1,* 443-490.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 0956797615616374.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271-9. doi:10.1080/17470218.2012.711335

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, *349*(6251), 1-8. doi:10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531-6. doi:10.1177/1745691612463401

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301-8. doi:10.3758/s13423-014-0595-4

Rouder, J. N., & Morey, R. D. (2011). A bayes factor meta-analysis of bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682-9. doi:10.3758/s13423-011-0088-7

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*, 1359-1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science: A Journal of the American Psychological Society / APS*, *26*(5), 559-69. doi:10.1177/0956797614567341

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*(4), 361-377.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology. General*, *143*(4), 1457-75. doi:10.1037/a0036731

Wassersman, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician, 70,* 129-133.

Author Note

Correspondence should be addressed to Peter Dixon, Dept. of Psychology, Univ. of Alberta, Edmonton, AB Canada T6G 2E9, peter.dixon@ualberta.ca.

Author Contributions

P. Dixon developed the method described in the paper and analyzed the data. S. Glover collaborated on the interpretation and the writing.

Figure Captions

Figure 1. Likelihood ratios under the null and replication models for three different obtained effects. The solid curve represents the expected distribution of effect sizes under the replication model, while the dotted curve represents the same under the null model.

Figure 2. Results of applying the likelihood-based approach to studies in the Reproducibility Project (Open Science Collaboration, 2015). In the top panel, gray areas indicate the frequency of ΔAIC (difference in AIC values; see Equation 2.10 in the supplementary materials), and dotted vertical lines indicate the criteria of ±3 ΔAIC. The dark gray in the center panel depict the subset of studies for which the effect of interest was significantly greater than 0. The dark gray areas in the bottom panel depict the subset for which the effect was significantly smaller than d_{33} (Simonsohn, 2015).

Figure 3. Relationship between log Bayes factor and AIC difference when applied to data from the Reproducibility Project (Open Science Collaboration, 2015). For clarity, four studies with an AIC difference greater than 20 are not shown.