Assessing Evidence for Replication

Peter Dixon

University of Alberta

and

Scott Glover

Royal Holloway University of London

Abstract

Recent discussion of the replication crisis has largely neglected a fundamental issue, namely how replication attempts are best evaluated. We develop a "good-faith" 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 that the researchers can reasonably be assumed to expect. A likelihood ratio is then calculated to contrast the match of two models to the data from the replication attempt: A model based on the anticipated effect size, and a model in which the effect is zero. When applied to data from the Replication Project (Open Science Collaboration, 2015), the procedure indicates that as many as 42.8% of the results failed to replicate.

Assessing Evidence for Replication

The Replication Crisis

There has been a great deal of concern expressed recently regarding the “replication crisis” in psychology (e.g., Lindsay, 2015; Pashler & Harris, 2012), in which a potentially large number of published results may be difficult to replicate. Replication problems have been ascribed to a number of factors, including data analysis strategies that inflate the Type I error rate (e.g., Simmons, Nelson, & Simonsohn, 2011), publication practices (e.g., de Bruin, Treccani, & Della Sala, 2015), and inherent problems with significance testing (e.g., Masicampo & Lalande, 2012). Any or all of these issues may indeed contribute to a failure to replicate, but an equally important question revolves around what actually counts as evidence for or against replication. In fact, it seems critical to have a solid statistical foundation for deciding whether a replication has been a success or failure before determining solutions for issues related to improving the reproducibility of results.

In the present paper, we first describe what we consider to be an appropriate framing of the replication question. We follow this by briefly reviewing some of the approaches to assessing replication, all of which seem to have shortcomings given our framing. We then develop what we refer to as a “good-faith” approach based on assumptions about the competence of the original researchers. Finally, as an illustration of the technique, we apply it to data from the Reproducibility Project (Open Science Collaboration, 2015).

What is Replication for?

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 a theoretically interesting effect must be of a certain magnitude regardless of statistical significance. We assume that published papers will generally provide reasonable evidence for theoretically interesting effects given that such evidence is a central criterion on which publication depends. And of course, published work is what is normally targeted for replication attempts.

From the perspective of the field and for the advancement of scientific knowledge, we normally would not care whether a replication produces precisely the same result as the original study; we care whether the replication evidence supports the same *interpretation* (that there is a theoretically interesting effect). However, the magnitude of a theoretically interesting effect can be difficult to determine. Although researchers may have an intuitive knowledge of how large an effect should be in order to be interesting, it is rarely discussed in research reports. The technique developed below provides one way to estimate the size of a theoretically interesting effect by examining the form of the original study. In essence, it is a way of gauging the researchers' expectations about effect size from the design of the study that they ran. Thus, the first critical aspect of a replication attempt is that it asks the question: Does the evidence from the replication support the existence of a theoretically interesting effect or not?

A second critical aspect of assessing evidence for replication is that one's concerns are symmetrical: We wish to be able to identify both when the evidence is in favor of replication and when it is against replication. If one can gauge the magnitude of a theoretically interesting effect, the replication question can be posed in this symmetrical fashion. The benefit of constructing such a symmetrical question is that it is straightforward to address statistically. That is, we can ask: Does the replication evidence support the existence of either a theoretically interesting effect, or does it support a null effect?

In contrast to posing the question symmetrically, extant replication techniques generally pose the question asymmetrically: Either they are designed to find evidence in favor of replication or they are designed to find evidence against it. The consequence of addressing the matter in this asymmetric fashion is that these techniques can only ever address one of the two questions of interest. Moreover, they are often silent on the issue of whether a replication attempt demonstrates a theoretically interesting effect or whether the effect is merely statistically significant. Below, we discuss several of these extant techniques and describe weaknesses that make them unappealing to those wishing to evaluate a replication attempt. Following this discussion, we introduce our "good-faith" approach and describe how it addresses both of the critical issues we described above: the issue of how to define a theoretically interesting effect, and whether the evidence from the replication attempt is either for or against it.

Existing Methods of Assessing Replication

Comparing Patterns of Significance. For many researchers, the most intuitive approach to deciding whether results replicate the original is to compare patterns of significance. Finding a significant result where was one found before would be construed as a replication, whereas failing to find such an effect would count as a failure to replicate. However, this method has a number of defects. An obvious one is that two results may be very similar, with largely overlapping confidence intervals, yet one is significant and the other is not. For example, if the original study obtained a significant result with a p value only a little below a criterion of .05, then a replication attempt with only a slightly smaller effect size might easily be nonsignificant. Under such circumstances, the two means would be very similar and the confidence intervals for the two results would largely overlap. This situation is illustrated in Panel 1 of Figure 1. Even though the results of the two studies are on opposite sides of the threshold for significance, their data are nearly indistinguishable, and it makes little sense to describe the result as a failure to replicate.

A related, but perhaps less well recognized, issue is that two results might both be significant but quite different. This is illustrated in Panel 2 of Figure 1. In this case, a significant result is obtained both in the original study and in the replication attempt. However, the effect in the replication attempt is much larger than that obtained in the original, and there is no overlap in the confidence intervals for the two effects. Thus, even though both have significant effects, there is clearly a large difference in the two results, and it would be misleading to describe the second as a successful replication.

Both of these problems reflect the fact that null hypothesis significance testing entails an arbitrary distinction between “significant” and “nonsignificant” results and requires that researchers behave differently in situations in which significant or insignificant results are obtained (cf. Gigerenzer, 2004). Dixon (2003) has described the problems that arise from this aspect of significance testing in factorial designs, where there can be little relationship between patterns of means and patterns of significance. However, the problem is also apparent in simple situations in which a single effect is being compared across two studies.

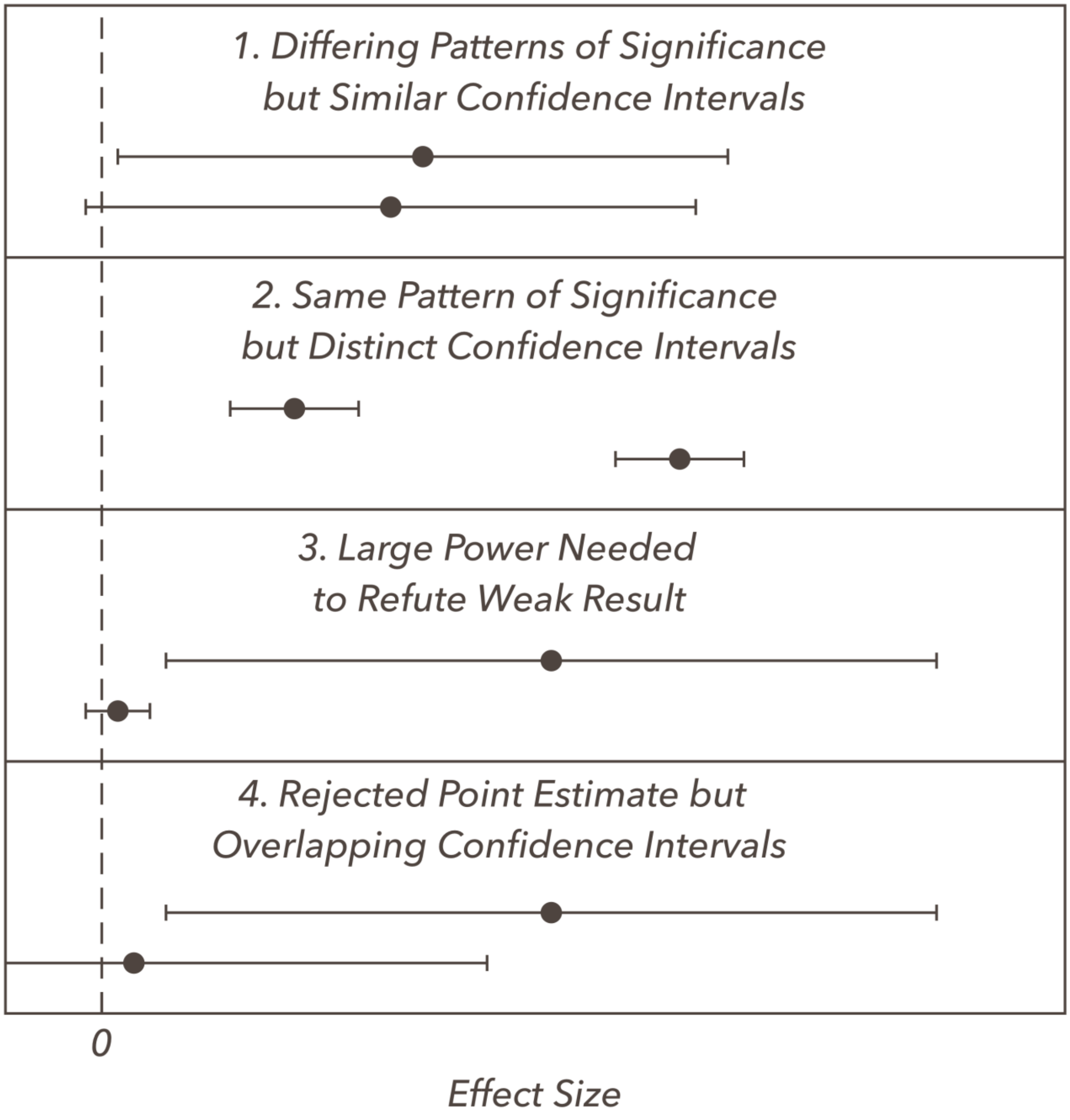
1. Differing Patterns of Significance but Similar Confidence Intervals
2. Same Pattern of Significance but Distinct Confidence Intervals
3. Large Power Needed to Refute Weak Result
4. Rejected Point Estimate but Overlapping Confidence Intervals
0
Effect Size

Figure 1. Possible relationships between effects found in a study and a replication attempt.

Testing for Differing Results. A somewhat more sophisticated attempt to compare two results might be to examine the confidence intervals of their respective effects. Essentially, if the intervals overlap, one would conclude that they are consistent with one another (and hence that the original study is replicated), and if the intervals don't overlap, or perhaps overlap only minimally, one would conclude that there was a failure to replicate. This procedure is tantamount to performing a significance test to see if the effects in the two studies differ.

Such an approach is problematic for two reasons: First, given the usual tenets of significance testing, one should not in principle draw any conclusions from a failure to find a significant difference between the two experiments. Consequently, there would be no procedure for finding evidence *for* replication; testing for a significant difference can only yield evidence *against* replication. Second, it may be very difficult to identify evidence for a failure to replicate if the original effect was small. Consider the situation in Panel 3 of Figure 1. In this scenario, the initial result has a confidence interval for the effect that extends nearly to (but does not include) zero. If the actual effect really is zero, then one might need to have a great deal of power in order to have a confidence interval centered at zero that does not overlap with that of the original study. Thus, according to this criterion, a very large sample size would be needed to demonstrate that a weak effect cannot be replicated.

A related approach is to test whether the results from the replication attempt are significantly smaller than the effect size obtained originally, as illustrated in Panel 4 of Figure 1. However, it is quite possible to obtain a significantly smaller effect even though the confidence intervals from the two studies overlap substantially (as shown in the figure). In other words, there could be a large range of effect sizes that are entirely consistent with both experiments.

Combining Evidence. Another approach is to combine the evidence from the original study and the replication attempt using meta-analysis or related techniques. One might then examine the combined evidence for the effect. If the original study found a weak (but significant) effect and the replication attempt found a weak (but nonsignificant) result, the combined evidence might be moderately in favor of the effect. However, combining this approach with significance testing makes it difficult to find evidence against replication because a confidence interval that includes zero does not imply that the effect is exactly zero. Moreover, in some cases, the motivation for attempting to replicate may be that there are concerns about the manner in which the original study was conducted or analyzed. Thus, it might not be appropriate to combine the results of the two studies as if they were independent samples from the same population.

Bayesian Approaches. Bayesian approaches to hypothesis testing provide several advantages over significance testing with respect to replication. For example, they are generally immune to problems involved in using optional stopping principles, so that data collection in a replication attempt can proceed until the evidence is clearly for or against replication (cf. Rouder, 2014). Nevertheless, many of the approaches that have been previously proposed are conceptually related to the ideas proposed in the context of significance testing. For example, one may use Bayesian hypothesis testing to assess whether there is evidence against the null hypothesis in a replication attempt, similar to testing whether the replication attempt produced a significant effect. Alternatively, one may compare the effect sizes in the original study and replication attempt using Bayesian hypothesis testing (Bayarri & Mayoral, 2002), parallel to testing whether two effect sizes are significantly different. A variation on this approach was proposed by Verhagen and Wagenmakers (2014) in which the effect size in the replication attempt is compared to the posterior distribution derived from the original result. Finally, it has been suggested that the original and replication attempts be combined to assess whether the evidence on aggregate favors the null or alternative hypothesis (e.g., Rouder & Morey, 2011), similar to evidence combining done in the context of significance testing. The conceptual similarities between these approaches and those devised in the context of significance testing suggests that such Bayesian replication methods may suffer from the same types of drawbacks outlined above. For example, regardless of whether one uses traditional NHST or Bayesian analysis methods, it is still the case that a powerful study would be necessary to provide evidence against a previously obtained weak result.

"Small Telescopes." Simonsohn (2015) described an interesting solution to some of the problems with previous approaches. Rather than assessing evidence for or against the 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. In particular, he starts by defining a "small effect" as an effect that could be found 20% of the time given the sample size used in the original study (described as "d_{20}"). Then, one conducts an analysis to see if the effect obtained in the replication attempt is significantly smaller than d_{20}. If the hypothesis is rejected, one could conclude that the original result fails to replicate because the effect must be smaller than what the original experiment could reasonably be expected to find. In other words, 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, but it might still require a relatively large sample to find evidence against a weak effect. Further, the procedure is set up essentially as a means to provide evidence that would discredit the original study, and a failure to reject the null hypothesis in this case provides only weak evidence *for* replication. In this sense, we regard it as a "bad-faith" approach in which one begins with the hypothesis that the original study was flawed and seeks evidence to support that hypothesis. Moreover, because it is based on null hypothesis significance testing, it is susceptible to some of the same issues that might have contributed to problems with the original study, such as optimal stopping and file-drawer issues.

A Good-Faith Approach

In contrast to the small-telescopes idea, we propose a "good-faith" approach to assessing evidence for replication. Although similar in some respects to the Simonsohn approach, it is based on the assumption that the original researchers understood the phenomena they were investigating and designed a suitably powerful study. We then estimate how large an effect the original researchers might have been expecting given the design they used and presume that this effect is large enough to be theoretically interesting. Based on this estimate of the expected effect size, we can measure the relative evidence for two alternative interpretations of the data from the replication attempt: 1) The result is consistent with the estimate based on the original study design (i.e., that the anticipated result was replicated); or 2) The effect is zero. Evidence for or against replication is the evidence in favor of either one or the other interpretation.

In the present development, we build on the approach to assessing evidence described by Glover and Dixon (2004). They suggested using an "adjusted" likelihood ratio to describe the evidence for one interpretation relative to another. The likelihood ratio is the likelihood of the data given one model (and its best-fitting parameters) divided by the likelihood of the data given another model (and its best fitting parameters). Such a ratio will nearly always favor the model with more parameters because such a model is more flexible. One way in which one can compensate for this additional flexibility is to adjust the likelihood ratio based on the Akaike (1973) Information Criterion. Such an adjusted likelihood ratio is:

$$\lambda_{adj} = e^{k_1 - k_2} \frac{L_1(X|\hat{\theta}_1)}{L_2(X|\hat{\theta}_2)} \tag{1}$$

where *X* is the vector of observations, $\hat{\theta}_1$ and $\hat{\theta}_2$ are the vectors of parameter estimates, L_1 and L_2 are the likelihoods under the two models, and k_1 and k_2 are the number of parameters. 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.” Often, one is interested in comparing a model in which there is an effect of some experimental factor to a null model in which there is no such effect, and one may assume that the data are normally distributed. In such cases, the adjusted likelihood ratio is:

$$\lambda_{adj} = e^{-k}(f^2 + 1)^{\frac{n}{2}} \tag{2}$$

where f^2 is a measure of effect size (Cohen, 1988), *k* is the effect degrees of freedom, and *n* is the number of independent observations.

In order to develop an index of evidence for replication, we work backwards from the design of the original study. We use the term “anticipated evidence” to refer to the adjusted likelihood ratio one would expect given a particular effect size and sample size. The good-faith approach to replication evidence is based on the view that the original researchers planned their experiment so that the anticipated evidence was reasonably strong. We argue that a plausible minimum value for such anticipated evidence would be 8 (corresponding to power of about .7). Then, from Equation 2, we can solve for the anticipated effect size as a function of *n*:

$$f_{ae}^2 = \left(8e^k\right)^{\frac{2}{n}} - 1 \tag{3}$$

Now, assume that in a replication attempt, an effect size of f^2_{obt} was observed. The evidence for replication is a likelihood ratio comparing two models of this observed effect size: a null model that assumes that the effect size is 0 and an "anticipated effect" model that assumes that the effect size is f^2_{ae}. This likelihood ratio is:

$$\lambda_{rep} = e^{-(k-1)} \left[\frac{f^2_{obt} + 1}{(f_{ae} - f_{obt})^2 + 1} \right]^{\frac{n}{2}} \tag{4}$$

In interpreting the results of applying Equation 4, very large values would provide clear evidence for replication, very small values would indicate evidence against replication, and values near 1 would be indeterminate. The numerator in this formulation corresponds to the error under the null model: Essentially, any obtained effect must be counted as error. The denominator corresponds to the error under the anticipated effect model: When the obtained effect is smaller than that anticipated, the difference must count as error. Of course, it is possible that the observed effect in the replication attempt is actually larger than the anticipated effect, but in this case, the observed effect will be larger still than an effect of zero, and the likelihood ratio will strongly favor replication. One may also express the evidence as the difference in AIC values for the two models, effectively changing the likelihood ratio to a log scale:

$$\Delta AIC = n \ln \left[f^2_{obt} + 1 \right] - n \ln \left[(f_{ae} - f_{obt})^2 + 1 \right] - (k-1) \tag{5}$$

As before, large positive values provide evidence for replication, large negative values provide evidence against replication, and values near zero would be indeterminate.

In the special case where the experiment involves comparing two conditions, the AIC adjustment represented by *k* in Equations 4 and 5 drops out. In this case, the two models being compared have the same number of parameters, and the likelihood ratio contrasts two point estimates of the effect size, 0 and d_{ae} (where $d_{ae} = 2f_{ae}$). Thus, for the comparison of two conditions, Equation 4 would not depend on any particular approach to measuring model complexity and, for example, a Bayesian analysis using BIC would yield the same result. Equation 5 could then be described either as a difference in AIC values or a difference in BIC values.

As a numerical example, suppose that the original researcher used a design with two independent groups of 20 subjects each, and found a significant effect of $F(1, 38) = 5.08$. This corresponds to an effect size of $f^2 = 0.1337$. If this result were described using an adjusted likelihood ratio following the approach of Glover and Dixon (2004), we would obtain:

$$\lambda_{adj} = e^{-k}\left(f^2+1\right)^{\frac{n}{2}} = e^{-1}\left(0.1337+1\right)^{40/2} = 4.53$$

This provides some evidence for a difference between groups but is rather smaller than the minimum anticipated evidence of 8 that we noted earlier. One might conclude either that the researcher was either wrong about the size of the effect being investigated or that the observed effect was relatively small just by chance. In any event, what is important for the present approach to replication is the size of the study, consisting of 20 subjects in two groups. From Equation 3, the anticipated effect size is inferred to be:

$$f_{ae}^2 = \left(8e^k\right)^{\frac{2}{n}} - 1 = (8e)^{2/40} - 1 = 0.1664$$

Suppose a second researcher attempted to replicate the result using a somewhat larger design with two groups of 30 subjects. If the observed effect size in the replication attempt was $f^2 = 0.1220$, by Equation 4 the evidence for replication would be:

$$\lambda_{rep} = e^{-(k-1)}\left[\frac{f_{obt}^2+1}{(f_{ae}-f_{obt})^2+1}\right]^{\frac{n}{2}} = \left[\frac{0.1220+1}{(\sqrt{0.1664}-\sqrt{0.1220})^2+1}\right]^{60/2} = 28.51$$

This provides clear evidence for replication. Alternatively, if the obtained effect size was half that size, $f^2 = 0.0642$, the evidence for replication would be:

$$\lambda_{rep} = e^{-(k-1)} \left[\frac{f_{obt}^2 + 1}{(f_{ae} - f_{obt})^2 + 1}\right]^{\frac{n}{2}} = \left[\frac{0.0642 + 1}{(\sqrt{0.1664} - \sqrt{0.0642})^2 + 1}\right]^{60/2} = 3.18$$

This provides some support for replication. Note, however, that this effect size corresponds to an $F(1, 58) = 3.72$, which would not quite be statistically significant. Finally, if only a small effect of 0.0161 is observed, the replication evidence would be:

$$\lambda_{rep} = e^{-(k-1)} \left[\frac{f_{obt}^2 + 1}{(f_{ae} - f_{obt})^2 + 1}\right]^{\frac{n}{2}} = \left[\frac{0.0161 + 1}{(\sqrt{0.1664} - \sqrt{0.0161})^2 + 1}\right]^{60/2} = 0.16$$

Or, inversely, 1/0.16 = 6.06 in favor of failing to replicate.

Although this technique provides a straightforward way to describe the evidence for and against replication, “failure to replicate” has a somewhat specialized meaning in this context: It means that the effect is smaller than what one might infer the original authors expected and is better described as 0. Even if there is substantial evidence against replication in this sense, it is still possible that the effect is nonzero but small. Indeed, a small effect that is difficult to detect is always a possibility on any approach to replication. If there is substantial evidence against replication using this procedure, it might mean that more consideration needs to be given to the question of how large an interesting effect would be, and perhaps more powerful studies would need to be designed to detect such an effect.

Application to the Reproducibility Project

As an illustration of the good-faith 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 can be replicated. 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. This resulted in a total of 84 pairs of studies. For each pair, we calculated the anticipated effect size for the original design using Equation 3, and the replication effect size from the reported test statistic (i.e., $f^2 = (df_1 / df_2)F$ or $f^2 = t^2 / df$). Equation 5 was then used to calculate the evidence (expressed as the difference in AIC values) for or against replication.

The results are shown in Figure 2. A value of 3 for ΔAIC might be considered “large” in this case. (For example, in some prototypical hypothesis testing situations, an obtained *p* value of .05 corresponds to a ΔAIC of 2.2.) By this criterion, 35.7% of the results demonstrated clear evidence for replication, and 42.8% demonstrated clear evidence against replication. The remaining 11.5% of cases were indeterminate. Of course, as noted above, it is quite possible that many of the failures to replicate (and many of the indeterminate results) represent real effects that are smaller than what one might infer the original authors expected.

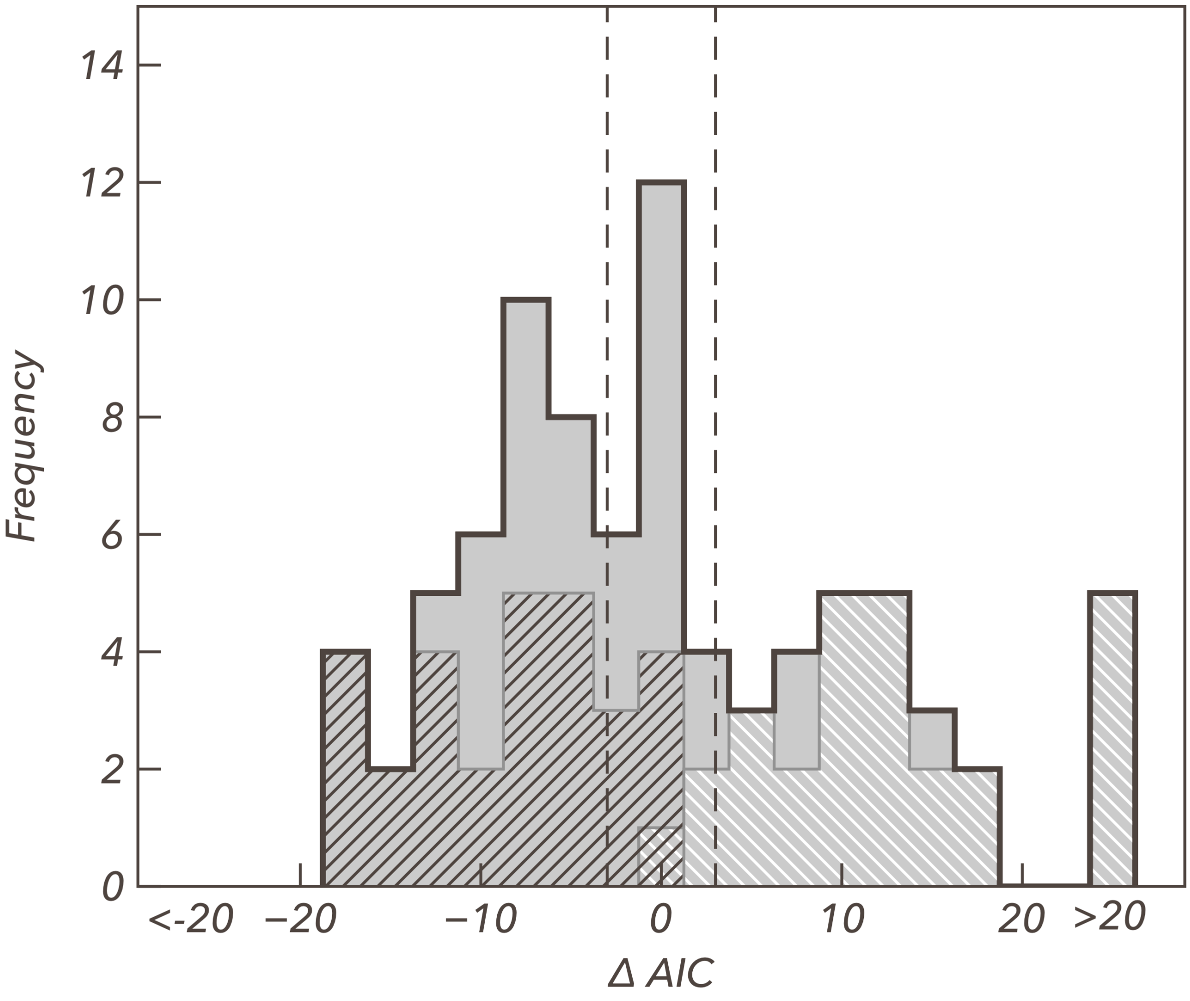

Figure 2. Results of applying the “good-faith” approach to studies in the Reproducibility Project (Open Science Collaboration, 2015). Gray areas indicate the frequency of ΔAIC (difference in AIC values; see Equation 5), and dotted vertical lines indicate the criteria of a -3 and +3 ΔAIC. Black hatched areas on the left indicate replication effect sizes significantly smaller than that in the original study; white hatched areas on the right indicate significant replication effects.

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 white hatched areas on the right in Figure 2. As can be seen, nearly all of the cases in which the good-faith assessment yielded clear evidence for replication were also statistically significant by this criterion. 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 actually closer to zero than to 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 does not always correspond to evidence for a failure to replicate using our good-faith approach.

In the second comparison, the effect size of the replication attempts was compared to those in the original studies directly. As an approximate index of whether the effect size in the replication attempt was smaller than that in the original study, both effect sizes were expressed as *r* and then transformed into *Z* scores using the Fisher transform. The difference in these *Z* scores would then be approximately normally distributed with a standard deviation, $\sqrt{1/(N_o - 3) + 1/(N_r - 3)}$. The black hatched areas on the left of Figure 2 represent those studies for which this difference was larger than the 95th percentile. As one might expect, when there was difference this large between the confidence intervals from the original and the replication study, there was generally clear evidence against replication. However, it is also apparent from Figure 2 that there were quite a number of studies for which there was substantial evidence against replication even though there was not compelling evidence for a difference in effect sizes. This follows from the argument made in the introduction that it can difficult to find a difference between a replication attempt and an original result with a confidence interval that extends nearly to zero. In such instances, the present technique is useful in demonstrating that the replication effect is smaller than what the original researchers might have expected. There are also a few cases in which there was a significant difference in effect sizes but not strong evidence against replication. This can happen if the effect size in the original study was substantially larger than one might expect based on the design, but the replication suggests an effect size more in keeping with the original design.

It is notable that using our good-faith 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 a more precise index of the problem. Although we do not have a simple interpretation of the nature of the replication problem, two issues occur to us. One is the extent to which the replication attempt matched the methods of the original study. This might be a problem in some portion of the studies; for example, the original authors approved the replication attempt in only 67.8% of the cases. When we consider only these replication attempts, the percentage of failures to replicate by our index is somewhat lower, 38.6%. Another issue is the strength of the evidence in the original study. For example, it is common practice in psychology to regard a *p* value of .05 as "significant" even though the evidence provided by such results is fairly weak. At least part of the lack of reproducibility may thus be a willingness to accept weak evidence for a conclusion. Indeed, the correlation between evidence for replication (as calculated here) and the adjusted likelihood ratio calculated from the original result is .677. A similar relationship was also noted by the Reproducibility Project (2015). However, in our case, evidence for replication is independent of the results actually obtained in the original study because our estimate is based only on the original study size. One part of the solution to the replication problem may thus be to insist on more compelling evidence.

Concluding Comments

Our good-faith approach to replication has several clear advantages over previous methods. First, it makes the (in our view) reasonable assumption that the original researchers competently designed their study to have sufficient power to detect a theoretically interesting effect. It then infers the expected effect size based on their design. Second, 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 generally can only answer one or the other side of the question. Third, 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.

Although we have developed and applied this technique in terms of adjusted likelihood ratios, the same concepts could be used regardless of one's approach to hypothesis testing. In particular, using the design of the original experiment to perform a good-faith assessment of the original research aims 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, the conclusions would typically be quite similar.

Understanding why results cannot be reproduced is a critical issue in psychology and other sciences, but assessing reproducibility is equally crucial, for without an accurate evaluation of the problem one cannot formulate suitable solutions. Our approach is to focus on the central question, “Does the data provide evidence for a theoretically interesting effect?”, and to frame this question in a symmetrical fashion. This “good faith” approach provides a graded and more principled means of assessing replication than many extant methods.

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.

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

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.

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.

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

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

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 (2006)*, *65*(11), 2271-9. doi: 10.1080/17470218.2012.711335

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

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. Possible relationships between effects found in a study and a replication attempt.

Figure 2. Results of applying the "good-faith" approach to studies in the Reproducibility Project (Open Science Collaboration, 2015). Gray areas indicate the frequency of ΔAIC (difference in AIC values; see Equation 5), and dotted vertical lines indicate the criteria of a -3 and +3 ΔAIC. Black hatched areas on the left indicate replication effect sizes significantly smaller than that in the original study; white hatched areas on the right indicate significant replication effects.