Understanding Replication

10

Contents

What You Will Learn in This Chapter

This chapter uses the power analyses from Chap. 7 and the meta-analytic methods from Chap. 9 to identify improper statistical analyses in published findings. The basic idea is simple. Unless the power is very high, we know that even real effects will not always produce significant outcomes simply due to random sampling. If power is only moderate but all studies are significant, the reported results seem too good to be true. Our considerations have a crucial implication: replication cannot be the final arbiter for science when hypothesis testing is used, unless experimental power is very high. Chapter 11 shows how such results can be produced even when scientists are trying to do everything properly.

10.1 The Replication Crisis

Across all sciences, replication is considered to be the "gold standard" for demonstrating important findings. Should a colleague happen to doubt the veracity of your empirical claim a surefire way to shut him down is to demonstrate that the effect can be consistently reproduced. The demonstration is especially effective if an independent lab replicates the effect. Along similar lines, if an independent lab reports that an effect cannot be replicated,

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_10

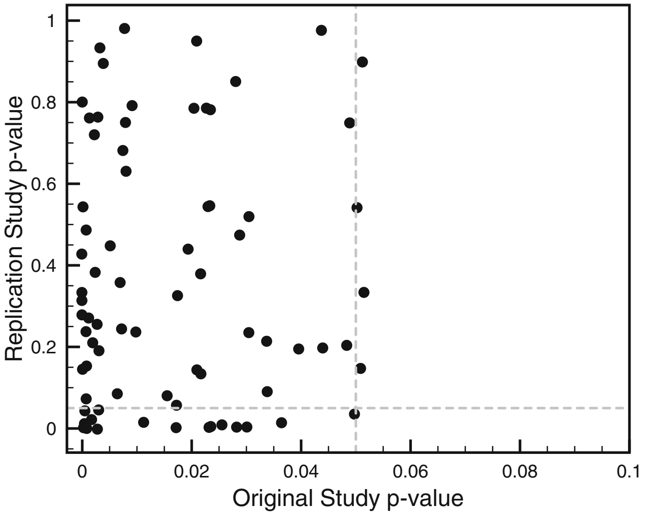

Fig. 10.1 Each point corresponds to a pair of p-values for an original study and its replication. While almost all original studies produced $p < 0.05$, very few replication studies produced such a small p-value. The figure is reproduced from Open Science Collaboration [1]. Please note the highly different scales of the x- and y-axes. The x-axis shows values in the range from 0.0 to 0.1, while the y-axis goes from 0.0 to 1.0. There is no obvious relationship between the p-values of the original and reproduction studies. A good result would have found that the p-values of the replication studies were smaller than 0.05, i.e., all black dots should be below the dashed horizontal line

there tends to be vigorous discussion about whether the correct procedures were followed and what the results mean. Successful replication is highly valued, and is taken as strong support for a scientific claim.

Unfortunately, many fields do not seem to be doing well with regard to replication. A group of psychologists, called the Open Science Collaboration [1], conducted replications of 97 studies that were published in three top journals. In their 2015 report, only 36% of the replication studies produced results consistent with the original studies. Each point in Fig. 10.1 plots the reported p-value for a replication study against the p-value for the original study. The dashed vertical line indicates the 0.05 criterion for the original studies, and almost all original studies reported a p-value below this criterion. This is not surprising because usually only significant results are published. The dashed horizontal line indicates the 0.05 for the replication studies, and almost all studies reported p-values above this criterion. The shocking result is that there hardly seems to be any relationship between the original study p-value and the replication study p-value. For example, some original studies reported p-values much smaller than 0.01 but the replication results yield p-values close to 1.0. Even worse, many of the replication studies had *larger* sample sizes than the original studies, and so should have produced smaller p-values, as we emphasized in Chap. 3 (Implication 2d).

Replication problems are not limited to psychology. In 2012, researchers at the biotech firm Amgen reported that they were unable to reproduce findings in 47 out of 53 landmark papers involving cancer research. There is an on-going effort by academic researchers to run replication studies similar to what was done by psychologists. The early results of that effort do not seem better than the replication results in psychology. For many people the lack of replication success in these studies indicates extremely serious problems that are sometimes referred to as a "replication crisis."

We agree that the problems are serious. However, we propose that rather than looking at new replication studies and wonder why they do not succeed, it is easier to look at the original published results and show that they never made sense.

Consider the following two phenomena that have been studied with multiple experiments.

- Phenomenon A: Nine of the ten experiments produced significant results, so it has a replication success rate of 0.9.
- Phenomenon B: Ten of the nineteen experiments produced significant results, so it has a replication success rate of 0.53.

If you follow the view that successful replication is a good guide for veracity, then the experimental outcomes definitely favor phenomenon A over phenomenon B. Neither phenomenon shows perfect replication, but we know from Chaps. 3 and 7 that not every experiment should work. Even so, phenomenon B only replicates about half the time, so we might even wonder whether the effect is real.

The problem with this interpretation is that phenomena A and B correspond to real investigations. Phenomenon A refers to what is known as precognition: the ability of people to get information from the future and use it in the present. A paper published in a top journal in 2011 reported that nine out of ten studies produced significant evidence for precognition. Despite the reported findings, very few scientists believe that precognition is a real effect; largely because its existence would undermine the very successful theory of general relativity. Thus, we are left to conclude that a high replication rate is not always sufficient to cause people to believe in the veracity of the effect.

Likewise, phenomenon B refers to what is known as the bystander effect: a tendency for people to not provide help to someone if there are other people around who could also provide help. Experiments on the bystander effect are rather difficult to run because one needs to have collaborators who pose as people needing help and other collaborators who pose as people who are around but not providing help. For this reason, these studies tend to use relatively small sample sizes. As a result, it is not uncommon for a study on the bystander effect to not produce a significant result. Even so, pretty much everyone believes that the bystander effect is a real phenomenon. Thus, we are left to conclude that a high replication rate is not always necessary to cause people to believe in the veracity of the effect.

We seem to be left with an odd situation. Scientists cite replication as a gold standard for judging the veracity of effects, but when faced with *actual* sets of experiments, replication seems neither sufficient nor necessary to establish veracity. It makes one wonder why scientists bother running experiments at all!

The way out of this odd situation requires a better understanding of statistics and replication. In the next subsection, we show that experiments should not always replicate, in particular, when effect and sample sizes are small. Replication success should reflect the estimated success probabilities of experiments. We should worry when experiments replicate too often.

10.2 Test for Excess Success (TES)

A set of experiments should succeed at a rate that follows the probability of success. Let us see whether that holds true for the precognition studies.

In the precognition study, each experiment was analyzed with a one-tailed, one-sample t-test. Table 10.1 lists the sample size for each experiment and the standardized effect size (Hedge's g). We use the meta-analytic techniques described in Chap. 9 to compute a pooled estimate of the standardized effect size. Doing so gives $g^* = 0.1855$. Doing the meta-analysis here is appropriate because the author of the studies used a similar analysis to provide partial support for his theoretical claim that precognition exists. Our pooled effect size, g^*, is our best estimate of the effect size, and we can use it to estimate the power of each individual experiment as described in Chap. 7. The last column in Table 10.1 shows the estimated power based on this meta-analytic effect size. Consistent with the observations about power made in Chap. 7, power values rise and fall with sample size. Experiment 9 ($n = 50$) is expected to have the smallest power (0.36) and Experiment 7 ($n = 200$) is expected to have the highest power (0.83). About half of the experiments (those with $n \approx 100$) have power values a bit above one half.

Table 10.1 Statistics for ten experiments that purported to find evidence for precognition

	Sample size (n)	Effect size (g)	Power
Exp. 1	100	0.249	0.578
Exp. 2	150	0.194	0.731
Exp. 3	97	0.248	0.567
Exp. 4	99	0.202	0.575
Exp. 5	100	0.221	0.578
Exp. 6a	150	0.146	0.731
Exp. 6b	150	0.144	0.731
Exp. 7	200	0.092	0.834
Exp. 8	100	0.191	0.578
Exp. 9	50	0.412	0.363

Suppose a scientist decided to replicate this set of ten experiments with the very same sample sizes as in the original report. If we accept the pooled effect size as a good measure of the precognition effect, then the expected number of successful outcomes in a set of ten experiments like these is the sum of the power values across the ten experiments. For example, if the power of each experiment is 1.0, then the number of significant results must be 10, the sum of the power values. For the studies in Table 10.1 the sum of the power values is 6.27. Hence, the experiments should have replicated 6.27 times. That expected degree of success is quite a bit lower than the 9 out of 10 success reported in the original investigation.

How likely is it to get 9 or 10 significant results for this effect? Doing something like a hypothesis test, we can estimate the probability of getting 9 or more successful outcomes from 10 experiments like these. We do not have to get exactly the 9 successes reported in the original report, any 9 out of 10 experiments will do. We compute the success probability by identifying all 11 combinations of experiments that demonstrate 9 or 10 successful outcomes. For each combination, we compute the probability of that particular result by multiplying the power of each successful experiment and the complement of power for each unsuccessful experiment. We then add up all those probabilities to get 0.058. That is, if the effect is real and similar to what was reported, a scientist doing a precise replication of the original ten experiments has only around a 6% chance of having the same degree of success as claimed in the original report. If replication success is supposed to guide our belief in the veracity of experimental results, this low rate seems like a serious problem.

Moreover, the low estimated replication rate begs the question of how the original author was able to produce such a high success rate. Given what we now know (from those studies) about the effect of precognition, it is very strange that those ten experiments were so successful. It is so strange that we can suspect that something went wrong in this set of experiments. We may never know exactly what happened in this set of experiments (even the original researcher might not know), but the burden of proof is on the researcher presenting the results. Perhaps there is a true precognition effect, but these studies do not provide good scientific evidence for it.

What if we apply the same kind of analysis to phenomenon B, where ten out of nineteen studies found statistically significant results for the bystander effect? Following the same basic approach, the pooled standardized effect size is -0.47, where the negative number indicates the presence of the bystander effect. That pooled effect size can be used to estimate the power for each of the nineteen experiments. The power varies from 0.2 to nearly 1.0 because several experiments had as few as 24 participants and one experiment had 2500 participants. Across all nineteen experiments, the sum of the power values is 10.77. Thus, we would expect to see around 11 significant results for experiments like these; and the nineteen experiments actually produced 10 significant results. Thus, the set of experimental results investigating the bystander effect seems believable, because the rate of success matches the estimated magnitudes of the effect and sample sizes of the experiments. The estimated probability of observing 10 or more significant results for studies like these is calculated to be 0.76.

10.3 Excess Success from Publication Bias

The previous subsection described the Test for Excess Success (TES), which examines whether the reported success rate of a set of experiments agrees with the estimated magnitude of the effect and the sample sizes of the experiments. If there is a big mismatch, then the TES suggests that there is a problem with the set of experiments, a problem with the analyses, or a problem with the theoretical claims based on the data/analyses. This subsection and the next use simulated experiments to show how it might happen that there is too much replication. This subsection considers the impact of publication bias: selective publishing of significant findings and suppression of non-significant findings.

Table 10.2 summarizes statistics from 20 simulated experiments that were each analyzed with a two-sample t-test. Each experiment had a simulated control group, for which there was no effect. For this group, scores were drawn from a normal distribution with a mean of zero and a standard deviation of one. For a simulated experimental group, scores were drawn from a normal distribution with a mean of 0.3 and a standard deviation of one. Hence, the population standardized effect size is $\delta = 0.3$. Sample sizes were the same for the two groups, $n_1 = n_2$. The sample sizes were drawn at random from a uniform distribution between 15 and 50.

The second column in Table 10.2 shows the t-value for each simulated experiment. The bolded t-values indicate statistical significance, as the p-values are less than the 0.05 criterion. There are five significant experiments. How does the success rate of five out twenty do when investigated with the TES? We can treat the simulated data in a way similar to the studies on precognition and the bystander effect. When we pool the effect sizes across all twenty experiments, we get $g^* = 0.303$. This estimated value is very close to the true value of 0.3, which simply demonstrates that meta-analysis works if all experiments are included in the analysis. We can use the pooled effect size to estimate power for each experiment, with the results reported in column 4 of Table 10.2. Summing these power values gives 4.2, and the probability of such experiments producing five or more significant outcomes is 0.42. There is no commonly agreed criterion for an appropriate success probability, but many people get concerned if the probability is less than 0.1. When both significant and non-significant experiments contribute to the analysis, the success rate tends to be consistent with the estimated power values. So far, so good.

Now suppose that a researcher practices a form of publication bias so that only the significant experiments (bolded t-values in Table 10.2) are published and available for further investigation. If we pool only the effect sizes for the five published experiments, we get $g^* = 0.607$, which is double the population effect size. This makes sense because those significant experiments must have a relatively large t-value. Since the effect size is a function of the t-value, these experiments must also have an unusually large estimated

Table 10.2 Statistics from twenty simulated experiments to investigate the effects of publication bias

$n_1 = n_2$	t	Effect size	Power from pooled ES	Power from biased ES
29	0.888	0.230	0.206	
25	1.380	0.384	0.183	
26	1.240	0.339	0.189	
15	0.887	0.315	0.126	
42	0.716	0.155	0.279	
37	1.960	0.451	0.251	
49	−0.447	−0.090	0.318	
17	1.853	0.621	0.138	
36	**2.036**	0.475	0.245	0.718
22	1.775	0.526	0.166	
39	1.263	0.283	0.262	
19	**3.048**	0.968	0.149	0.444
18	**2.065**	0.673	0.143	0.424
26	−1.553	−0.424	0.189	
38	−0.177	−0.040	0.257	
42	**2.803**	0.606	0.279	0.784
21	1.923	0.582	0.160	
40	**2.415**	0.535	0.268	0.764
22	1.786	0.529	0.166	
35	−0.421	−0.100	0.240	

Bolded t values indicate statistical significance ($p < 0.05$)

effect size. Hence, one impact of a publication bias is that the published studies can dramatically overestimate the magnitude of effects. Using the overestimated effect size to compute power for each experiment produces the values in the last column of Table 10.2. These values are dramatically larger than the true power values because they are based on a gross overestimate of the effect size. Nevertheless, the power values sum to 3.13, which indicates that out of five published experiments like these we would expect around three significant results. In reality, all five experiments produced significant results, and the probability that all five experiments would produce a significant result is the product of the power values, which is 0.081. For many people this is such a low probability (e.g., less than 0.1) that they would doubt the validity of the published results.

10.4 Excess Success from Optional Stopping

As mentioned in Chap. 4, a requirement for the t-test is that the sample sizes for the two groups are fixed before the experiment. In practice, however, it is very common for a sample to *not* have a fixed size. Consider the following situation. A scientist gathers data from two populations and ends up with $n_1 = n_2 = 10$ scores in each sample. The scientist runs a t-test and computes $p = 0.08$. This p-value does not fall below the 0.05 criterion that is used for statistical significance, but it looks promising. Oftentimes researchers in this situation decide to gather ten more scores, so that they now have $n_1 = n_2 = 20$ scores in each sample. Suppose that when the t-test is run on this larger sample it produces $p = 0.04$, which indicates statistical significance. This sounds good: more data gives a better answer. Unfortunately, this kind of procedure can dramatically inflate the Type I error rate. One problem is that this procedure involves multiple tests. Each test has some probability of producing a Type I error. As shown in Chap. 5, with multiple tests the probability of at least one of them making a Type I error is higher than the probability of a single test producing a Type I error.

The more serious problem with this procedure is that data collection is stopped once a desired result has been found. As additional observations are added to the original data set, a conclusion of significance may switch to non-significance, and vice-versa. If the decision to add data is tied to finding a significant result (e.g., no more data is collected once $p < 0.05$), then the data collection process is biased toward producing significant outcomes. This kind of procedure is called "optional stopping," and it increases the Type I error rate. An unscrupulous scientist who started with $n_1 = n_2 = 10$ and added one observation to each data set until getting a significant outcome ($p < 0.05$) or a maximum of $n_1 = n_2 = 50$ would have a Type I error rate over 20%.

It is important to recognize that the problem here is not with *adding* data but with *stopping* data collection because the Type I error rate refers to the *full* procedure. Thus, optional stopping is a problem *even* if the first data set happens to produce a significant result, but the scientist *would have* added more subjects to a non-significant data set. Importantly, if a researcher does not have a specific plan for data collection, then it is impossible to compute the Type I error rate. This is why the standard approach to hypothesis testing assumes a fixed sample size.

The TES is sensitive to a set of studies where researchers followed this kind of improper approach, and it is fruitful to look at simulated experiments to get some intuition on what happens. Table 10.3 summarizes statistics from 20 simulated experiments that were analyzed with a two-sample t-test. For both the control and experimental groups, the sample sizes n_1 and n_2 were the same. Scores were drawn from a normal distribution with a mean of zero and a standard deviation of one. Hence, the population effect size is $\delta = 0$; there is truly no effect here.

To simulate optional stopping, each sample started with $n_1 = n_2 = 15$ scores. A t-test was run on that data and if a significant result was found, the experiment was stopped and reported. If the t-test did not find a significant result, one more data point was sampled for each group and the t-test was repeated. This process continued up to a sample size of $n_1 = n_2 = 100$, where the result was reported.

Since the population effect equals zero, we would expect to get, on average, one significant outcome from twenty simulated experiments (see Chap. 5). The four bolded t-values in Table 10.3 indicate statistical significance, which is a much higher rate (20%)

Table 10.3 Statistics from twenty simulated experiments to investigate the effects of optional stopping

$n_1 = n_2$	t	Effect size	Power from pooled ES	Power from file drawer ES
19	**2.393**	0.760	0.053	0.227
100	0.774	0.109	0.066	
100	1.008	0.142	0.066	
63	**2.088**	0.370	0.060	0.611
100	0.587	0.083	0.066	
100	−1.381	−0.195	0.066	
100	−0.481	−0.068	0.066	
100	0.359	0.051	0.066	
100	−1.777	−0.250	0.066	
100	−0.563	−0.079	0.066	
100	1.013	0.143	0.066	
100	−0.012	−0.002	0.066	
46	**2.084**	0.431	0.057	0.480
100	0.973	0.137	0.066	
100	−0.954	−0.134	0.066	
100	−0.136	−0.019	0.066	
78	**2.052**	0.327	0.062	0.704
100	−0.289	−0.041	0.066	
100	1.579	0.222	0.066	
100	0.194	0.027	0.066	

Bolded t values indicate statistical significance ($p < 0.05$)

than the intended 5%. A simple computation, using the binomial distribution, shows that the probability of getting four or more significant experiments in a set of twenty is 0.016 when each experiment has a 5% chance of producing a significant result. All of the non-significant experiments in Table 10.3 have sample sizes of 100 (the maximum possible sample size) because that is the nature of the optional stopping procedure.

Computing the pooled effect size across all twenty experiments finds $g^* = 0.052$, which is very close to the population effect size of zero. Contrary to the effect of publication bias, optional stopping does not bias estimates of the effect size. Likewise, if we use that estimated effect size to calculate power for each experiment, we get values ranging from 0.053 to 0.066, which are all just above the 0.05 significance criterion because the estimated effect size is just larger than zero. Still, the reported results seem too good to be true. Adding up the power values for all twenty experiments gives just 1.28, so we would expect to find around one significant experiment among twenty experiments like these. The probability of experiments like these producing four or more significant outcomes is calculated from the power values as 0.036. This result (correctly) indicates some kind of problem in the set of experiments: the rate of success is larger than it should be.

The last column of Table 10.3 shows power values based on only the significant experiments in Table 10.3. Here, we suppose that the non-significant experiments were not published (publication bias). In that situation the TES analysis has to work with only the four reported significant experiments. The pooled effect size estimate is $g^* = 0.4$, which is dramatically larger than the true value of zero. As a result of this overestimate of the effect size, the power values for the four significant experiments are also dramatically overestimated. Nevertheless, adding up those four power values indicates that four experiments like these would be expected to produce around two significant outcomes. The probability of all four experiments producing significant outcomes is the product of the power values, which is 0.047. Again, this set of studies (correctly) seems problematic because the success rate is out of line with the estimated effect and the experiment sample sizes.

10.5 Excess Success and Theoretical Claims

The Test for Excess Success is able to identify situations where the reported rate of success does not match the experimental effect and sample sizes. An important point of this analysis is the definition of "success," which is always relative to some theoretical claim. As an example, suppose that a researcher runs ten independent experiments that each investigates a different topic (e.g., the Stroop effect, a memory experiment, differences in EEG alpha synchrony, epigenetic transfer of learned behavior, precognition, and other topics). Suppose that the first four experiments find a significant outcome but the other six experiments do not. Further suppose that the researcher imposes a publication bias and only publishes the four successful experimental results and does not publish the six null results found for the other studies. A TES analysis on the four published studies may (correctly) indicate evidence of publication bias, but this observation is fairly meaningless. The four experiments are unrelated to each other and are unrelated to any overarching theoretical claim. As such, all we can conclude is that there were other unsuccessful experiments that have not been reported, but the existence of such unsuccessful experiments tells us nothing about the veracity of the reported properties of the Stroop effect or performance in the memory experiment.

On the other hand, if the same researcher used the results of the very same four significant experiments to make some theoretical claim (e.g., a unified theory of the Stroop effect, memory, EEG alpha synchrony, and epigenetic transfer), then publication bias potentially undermines that theoretical claim. If a TES analysis indicates that the set of four studies suggests publication bias, then scientists should be skeptical about the corresponding theoretical claims that have been derived by the researcher.

Oftentimes researchers unintentionally make their theoretical conclusions seem too good to be true by having their theory be determined by the significance/non-significance of their tests. In this case the theory becomes nothing more than a coarse summary of what was measured in the experiment. Such a theory is almost certain to be chasing (some) noise in the experimental results and is almost surely not going to be fully supported by a new set of experiments.

Consistent with Chap. 3, Implication 3a, the conclusion of the TES analysis does not prove that there is no effect across a set of experiments; rather it indicates that the set of experiments does not make a convincing scientific argument.

Take Home Messages

1. If many similar experiments with low effect and sample size all lead to significant results: the data seem too good to be true.
2. Experiments should lead to significant results proportional to their power.
3. Publication bias and optional stopping can lead to strongly inflated Type I error rates.

Reference

1. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349. https://doi.org/10.1126/science.aac4716.