# Correspondence in Significance

## Model & Framework

One approach to assessing replication proceeds from standard meta-analytic models. Meta-analysis represents the results of an experiment by an effect parameter, denoted $\theta$. This "true" effect reflects what the researcher would observe if they had an infinite sample size. Effect parameters are typically on a standard scale, such as a standardized mean difference, Fisher-transformed correlation coefficient, or log-odds ratio.

In the context of replication studies, we have at least $k \geq 2$ such findings, so let $\theta_i$ describe the results of the $i$th study. For programs such as that of the Replication Project: Psychology (RPP) (Open Science Collaboration, 2015), this involves $\theta_1$ for the original study and $\theta_2$ for the replicate. However, we do not actually observe $\theta_1$ and $\theta_2$, and instead we estimate them by $T_1$ and $T_2$, and these estimates have variances $v_1$ and $v_2$. The estimates and their variances are what get reported in typical statistical analyses of experiments.

A common assumption in meta-analysis is that the $T_i$ are independent, unbiased, normally distributed, and have a known variance, that is:

$$T_i \overset{indep}{\sim} N(\theta_i, v_i)$$

This is true for effect sizes such as $z$-transformed correlations, and is a large sample approximation for other effect sizes such as standardized and raw mean differences and log-odds ratios. Further, in this paper, we assume that within-study hypothesis tests involve $H_0 : \theta_i = 0$, and are carried out at the $\alpha = 0.05$ level.

We argue that one useful definition of replication involves the similarity not of estimates $T_i$, but of actual true effects $\theta_i$. The $T_i$ might differ because of both differences in the underlying true effects, but also due to random chance. The same can be said for the resulting $p$-values—$p_1$ and $p_2$. The $\theta_i$ might differ, however, because studies sample participants from different populations, or due to differences in contexts and conditions between experiments. Thus, it may make sense to define replication in terms of differences in experiments rather than statistical noise. Thus, we offer $\theta_1 = \theta_2$ as a definition of replication worth testing.

## Correspondence in Statistical Significance

One metric used by the RPP to determine if studies successfully replicated was if the original and replicate study corresponded in sign and statistical significance. Findings were deemed to have replicated if both studies found significant effects in the same direction—i.e., both positive—or if both studies found null effects. Heuristically, this would seem logical, and indeed, Fisher (1935) noted that "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results."

However, it has long been noted that differences between significant and nonsignificant findings need not correspond to differences in underlying effect parameters (e.g., Gelman & Stern, 2005). Thus, it is entirely possible that $\theta_1 = \theta_2$, so that the studies replicate exactly, but only $T_1$ is statistically significant. Alternatively, it could happen that $\theta_1$ and $\theta_2$ are markedly different, but both studies involve statistically significant estimates.

To get a sense of the properties of this test, assume that $\theta_1 = \theta_2 = \theta$, or that the studies replicate exactly. If this were true, we would want this test to infrequently reject this null hypothesis; in statistical terms, we would want the type I error rate to be low. The null hypothesis would be rejected if $T_1$ is significant and $T_2$ is either not significant or significant and has an opposite direction (or visa versa). If both $T_1$ and $T_2$ are subject to two-tailed $\alpha = 0.05$ level tests, the type I error rate is given by

$$
\begin{aligned}
P[\text{Type I Error}] = &[1 - \Phi\left(1.96 - \theta/\sqrt{v_1}\right)]\Phi\left(1.96 - \theta/\sqrt{v_2}\right) \\
&+ \Phi\left(1.96 - \theta/\sqrt{v_1}\right)\left[1 - \Phi\left(1.96 - \theta/\sqrt{v_2}\right)\right] \\
&+ \Phi\left(-1.96 - \theta/\sqrt{v_1}\right)\left[1 - \Phi\left(-1.96 - \theta/\sqrt{v_2}\right)\right] \\
&+ [1 - \Phi\left(-1.96 - \theta/\sqrt{v_1}\right)]\Phi\left(-1.96 - \theta/\sqrt{v_2}\right)
\end{aligned}
$$

where $\Phi$ is the standard normal distribution function.

When $\theta_1 = \theta_2 = 0$, so that both studies involve null effects, then the type I error rate is 9.5%. However, when $\theta_1 = \theta_2 \neq 0$, the frequency of type I errors depends on the power of each study to detect a non-null effect. Figure 1 shows the false positive rate of the test as a function of the power of study 1, and the relative size of study 2 to study 1. For instance, if study 1 had 60% power to detect $\theta_1$ and study 2 had 60% power to detect $\theta_2$—so that the studies are the same size—then the false positive rate will be just under 50%!

In Figure 1, we see that the error rate is largest when the power of both studies is moderate. If both studies have high power, then it becomes more likely that both estimates will be significant; and if both studies have low power, then it becomes more likely that both estimates will be nonsignificant. However, the false positive rates in Figure 1 are nearly all above the nominal 5%. This is not necessarily an artifact of the scale of the graph. Indeed, if study 2 had an infinite sample size, so that $v_2 \to 0$ and $\theta_2/\sqrt{v_2} \to 0$, then the false positive rate of this test for replication converges to one minus the power of the initial study:

$$\text{Error rate } \to \beta_1$$

where $1 - \beta_1$ is the power of study 1. Thus, unless study 1 has at least 95% power, determining if studies replicated according to correspondence of sign and statistical significance can result in type I error rates well above 5%.

For reference, suppose all 100 of the RPP studies involved studies that replicated exactly, with effects that are nonzero ($\theta_1 = \theta_2 \neq 0$). If all studies were powered at 80%, then we would expect this procedure to falsely determine that 41 findings failed to replicate. If the power of the studies was lower, say 60%, then we would expect this procedure to falsely conclude that 48 of findings failed to replicate.

Conversely, if we assume $\theta_1 \neq \theta_2$, then we would want the procedure to determine that the findings do not replicate frequently—that is, we would want a low type II error rate. To conclude that the studies do replicate would require both findings to be significant and in the same direction (e.g., positive), or to both be nonsignificant. Under the model, the probability of concluding that the studies do replicate when $\theta_1 \neq \theta_2$ can be written as:

$$
\begin{aligned}
P[\text{Type II Error}] = &[1 - \Phi\left(1.96 - \theta_1/\sqrt{v_1}\right)][1 - \Phi\left(1.96 - \theta_2/\sqrt{v_2}\right)] \\
&+ [\Phi\left(1.96 - \theta_1/\sqrt{v_1}\right) - \Phi\left(-1.96 - \theta_1/\sqrt{v_1}\right)][\Phi\left(1.96 - \theta_2/\sqrt{v_2}\right) - \Phi\left(-1.96 - \theta_2/\sqrt{v_2}\right)] \\
&+ \Phi\left(-1.96 - \theta_1/\sqrt{v_1}\right)\Phi\left(-1.96 - \theta_2/\sqrt{v_2}\right)
\end{aligned}
$$

When one of the effect parameters, say $\theta_2$ is equal to zero, then this can be written in terms of the power of the other study:

$$P[\text{Type II Error}] = 0.025 + 0.925\beta_1$$

where $1 - \beta_1$ is the power of study 1. Thus, if study 1 has low power, so that $\beta_1$ is large, then this test would be likely to conclude that the studies successfully replicated. For instance, if study 1 has power 80%, then this test would conclude that the studies do replicate 21% of the time. However, if the power were lower, say 60%, then the test would make this error 39.5% of the time.

If both of the parameters are nonzero but the same direction (e.g., both positive), then the error rate will be large whenever both studies have high (or low) power. For instance, if $\theta_1$ is much larger than $\theta_2$, but both have power 80%, then this test will conclude that the studies replicate almost 68% of the time! Similarly, if

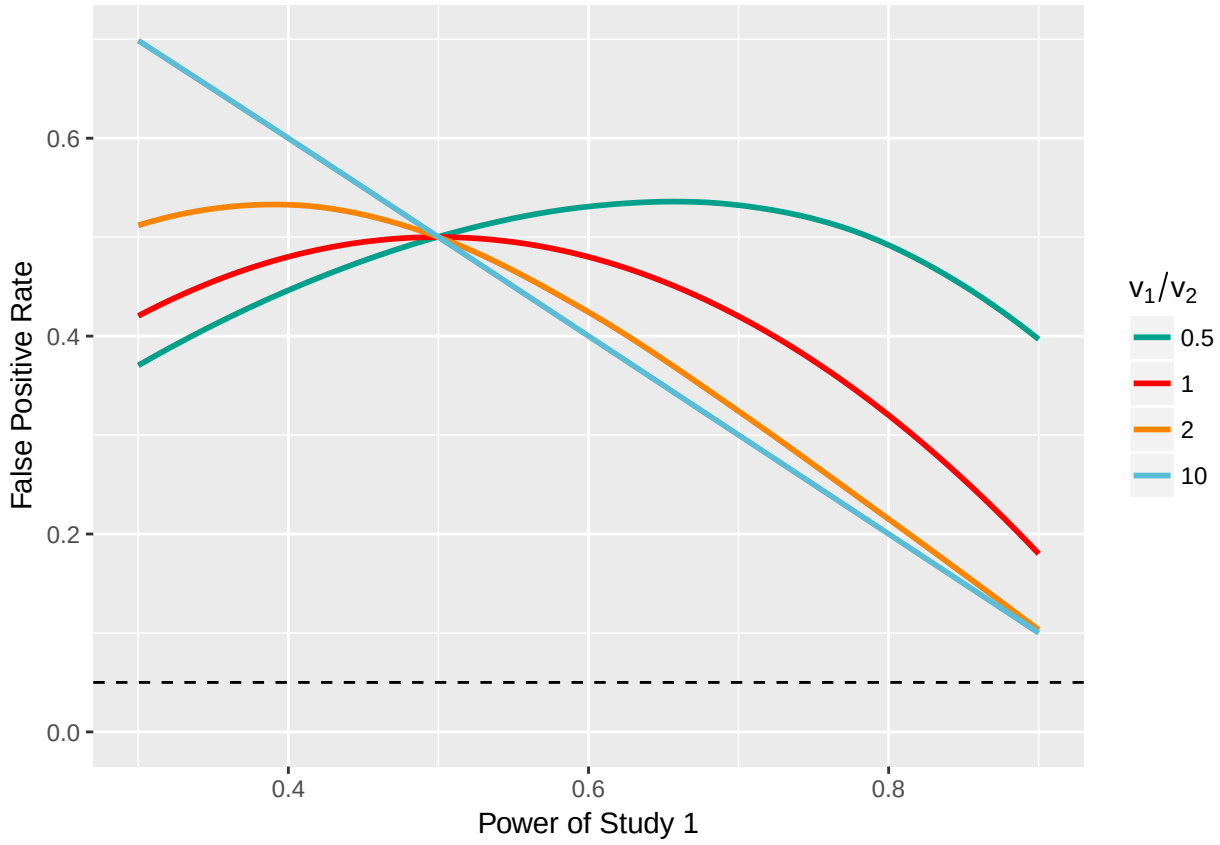


Figure 1: False Positive Rate of Test for Replication. This plot displays the false positive rate of determining replication according to correspondence of sign and statistical significance as a function of the size of study 1 and study 2. Each colored line corresponds to study 2 being a multiple of the size of study 1. The dashed line corresponds to a 5% false positive rate.

both studies have low power then there is a higher probability that both will return nonsignificant results, and hence we would conclude that the studies replicated despite $\theta_1 \neq \theta_2$. In fact, the only way for the false negative rate to decrease is for one study to have very high power, and the other to have very low power.

In summary, the properties of this metric will depend heavily on the power of each study to detect a non-null effect, but in a way that can be counterintuitive. If the studies replicate exactly, then this test will falsely conclude that they do not a large percentage of the time unless both either have very high power (e.g., >90%) or very low power (e.g., < 10%). Moreover, it will be difficult to find scenarios where the type I error rate is below 9.5%, and realistically could be much higher than that. Conversely, if the studies involve very different effect parameters, the test will often conclude that studies replicate. If effects are in the same direction, the power of this test will decrease in the sample size of each individual experiment. Only if one effect is positive and the other is null or negative will the error rate decrease with sample size. In other words, there are fairly plausible scenarios where collecting more data—having larger studies—can *increase* the type I or type II error rates of this test!