Assessing Replication via P-values

Sarah Peko-Spicer & Mena Whalen

June 19, 2018

Two of the methods proposed by the Open Science Collaboration (2015) for assessing replication via p-values are McNemar's test and the t-test for paired data. We discuss these tests in tandem as they are both extensions on the sign test and, as a result, have similar properties and agreement of significant results in the context of replication. For both tests, we consider k pairs of studies, each containing an original study and a replication study. As described earlier, the original studies are characterized by an estimate $T_{1j} \sim N(\theta_{1j}, v_{1j})$ for $j = 1, \ldots, k$. Similarly, the replication studies are characterized by an estimate $T_{2j} \sim N(\theta_{2j}, v_{2j})$ for $j = 1, \ldots, k$. Further, we let p_{1j} denote the p-value of the j^{th} original study and p_{2j} denote the p-value of the j^{th} replication study.

McNemar's test is used to determine if there is a difference between the proportion of statistically significant results among original studies and the proportion of statistically significant results among replication studies. Each p-value from an original or replication study can be categorized as either significant or non-significant, creating a dichotomous variable suitable for the use of this test. This information is generally summarized in a 2×2 contingency table as shown below where n_{11} denotes the number of study pairs in which both the original and replication findings are signficant, n_{10} denotes the number of study pairs in which the original finding is significant but the replication finding is not, n_{01} denotes the number of study pairs in which the original finding is not significant but the replication finding is, and n_{00} denotes the number of study pairs in which neither the original finding nor the replication finding is significant.

	Replication Finding	
Original Finding	Significant	Non-Significant
Significant	n_{11}	n_{10}
Non-Significant	n_{01}	n_{00}

The null hypothesis for McNemar's test states that the marginal probabilities for each outcome are the same. That is, $H_0 : p_{11} + p_{10} = p_{11} + p_{01}$ where the p's are the probabilities of occurrence in cells with the corresponding labels. The test statistic is given by $\chi^2 = \frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}}$. Under the null hypothesis, the test statistic follow a chi-squared distribution with one degree of freedom. The null hypothesis is reject for large values of the test statistic. The results of the OSC article are summarized in the contingency table below. 97% of original findings were significant compared to 36% of replication findings. The authors found a test statistic of $X^2 = 59.06$, rejected the null hypothesis, and concluded that the findings from the original studies did not replicate.

	Replication Finding	
Original Finding	Significant	Non-Significant
Significant	35	62
Non-Significant	1	2

Unlike McNemar's test, the t-test assesses replication by directly comparing original and replication p-values within a study pair. The null hypothesis for the t-test states that the original and replication studies have the same p-value (i.e. $H_0 : p_{1j} = p_{2j}$). The test statistic is given by $t = \frac{\bar{X}_{diff}}{SD_{diff}/\sqrt{k}}$ where $\bar{X}_{diff}$ is the mean of all differences between original and replication p-values (i.e. $\bar{X}_{diff} = \frac{1}{k}\sum_{j=1}^{k}(p_{1j} - p_{2j})$), SD_{diff} is the standard deviation of those differences from the mean, and k is the number of study pairs. A paired t-test is used in

this case since the original and replication studies in each study pair are conducting the same experiment, using the same methods, and drawing conclusions on the same phenomenon.

In the OSC article, the authors used the t-test on $k = 99$ studies. They omitted one study pair where the exact p-value was not available for the original or replication study. For the 99 available study pairs, the authors found a mean difference, $\bar{X}_{diff} = -0.274$, indicating that the replication p-values were larger than the original p-values. The test statistic was -8.207 ($p < 0.0001$). Thus, the authors concluded that the original and replication p-values are significantly different and the original and replication findings are significantly different.

It should be noted that both the t-test and McNemar's test are used to assess similarity between a *group* of original studies and a *group* of replication studies. We assume that an original and replication study within a study pair are conducting the same experiment, using the same methods, to draw conclusions about the same phenomenon. However, in the OSC data we observe some variation between study pairs on these factors. That is, the 100 original studies may be conducting different experiments, using different methods, to draw conclusions about different phenomena. On these grounds alone, the t-test and McNemar's test are not well suited to assess whether an original finding has been replicated unless all k study pairs are drawing conclusions about the same phenomenon. In the analysis that follows, we assume that this is, in fact, the case.

Analysis of McNemar's Test

Two properties you might use to assess the quality of a statistical test for the purpose of replication are its *false invalidation* and *false validation* rates. These rates are similar in nature to Type I and Type II error rates, but use a different null hypothesis as a point of reference. Assuming that the null hypothesis for replication is given by $H_0 : \theta_{1j} = \theta_{2j}$ and that the null hypotheses for the tests used in the OSC article are considered an "approximation" of this null, a false invalidation occurs when $\theta_{1j} = \theta_{2j}$ but the chosen statistical test rejects its null hypothesis. Similarly, a false validation occurs when $\theta_{1j} \neq \theta_{2j}$ but the chosen statistical test fails to reject its null hypothesis. To better understand these properties for McNemar's test, we rewrite the null hypothesis to reflect its relationship to the power of the original and replication studies. First, note that that the null hypothesis can be rewritten as

$$H_0 : P(p_{1j} \leq 0.05) = P(p_{2j} \leq 0.05)$$

In other words, under the null hypothesis, if we were to draw a study pair at random, the probability of obtaining a significant result in the original study is equal to the probability of obtaining a significant result in the replication study. Taking this observation into consideration and noting that $p_{ij} \leq 0.05 \equiv \frac{|T_{ij}|}{\sqrt{v_{ij}}} \geq 1.96$, we can reformulate the null hypothesis as follows.

$$H_0 : \sum_{j=1}^{k} P\Big(\frac{|T_{1j}|}{\sqrt{v_{1j}}} \geq 1.96|i\Big)P(i) = \sum_{j=1}^{k} P\Big(\frac{|T_{2j}|}{\sqrt{v_{2j}}} \geq 1.96|i\Big)P(i)$$

That is, we can think of McNemar's test as a test of difference in the average power rather than a difference in paired proportions. As such, a rejection of the null hypothesis is akin to concluding that there is a statistically significant difference between the average power of the original studies and the average power of the replication studies. However, this does not necessarily imply that the original findings were not replicated in the replication studies, if we take replication to be $\theta_{1j} = \theta_{2j}$. Similarly, a failure to reject the null hypothesis for McNemar's test does not necessarily imply that the findings did, in fact, replicate.

Assuming the true null hypothesis for replication is $H_0 : \theta_{1j} = \theta_{2j}$, a false invalidation occurs when the original and replication studies differ in their power to detect their common effect size. In this case, McNemar's test will almost always reject its null hypothesis leading to conclusions of irreplicability despite having identical true effects. The power of each of the $2k$ original and replication studies is determined by the true effect size

(θ_{ij}), the sample size (n_{ij}), and the population variance (σ^2). If the original and replication studies have the same true effect size and population variance, we can let their powers differ by increasing the sample size of the replication studies. As n_{2j} grows large, $\frac{|T_{2j}|}{\sqrt{v_{2j}}}$ will also grow large. If we assume that original and replication studies are independent then,

$$n_{10} = \sum_{j=1}^{k} \mathbf{1}\Big\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} \geq 1.96\Big\}\mathbf{1}\Big\{\frac{|T_{2j}|}{\sqrt{v_{2j}}} < 1.96\Big\} \to 0$$

$$n_{01} = \sum_{j=1}^{k} \mathbf{1}\Big\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} < 1.96\Big\}\mathbf{1}\Big\{\frac{|T_{2j}|}{\sqrt{v_{2j}}} \geq 1.96\Big\} \to \sum_{j=1}^{k} \mathbf{1}\Big\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} < 1.96\Big\}$$

This implies that the McNemar's test statistic $X^2 \to \sum_{j=1}^{k} \mathbf{1}\{\frac{|T_{1j}|}{\sqrt{v_{1j}}} < 1.96\}$ as the replication sample sizes grow large. Let's assume that all k original studies have 80% power to detect their true effect. Then, X^2 approaches $0.2k$ as the replication sample sizes increase. Recall that McNemar's test rejects if $X^2 > \chi^2_{[0.05,1]}$. In this instance, as long as $k > 2$, McNemar's test will always reject despite the true effect sizes being identical for all k study pairs.

A false validation occurs when the original and replication studies have different true effects (i.e. $\theta_{1j} \neq \theta_{2j}$) but have the same power to detect those effects (i.e. $(1 - \beta_{1j}) = (1 - \beta_{2j})$). The impact of the power of the original and replication studies on this test's false validation rate is easiest understood through a consideration of the test's power function. The power function for McNemar's test is given by

$$\Phi\Big\{\frac{(p_{10} - p_{01})\sqrt{k} - z_{1-\alpha/2}\sqrt{p_{10} + p_{01}}}{\sqrt{p_{10} + p_{01} - (p_{10} - p_{01})^2}}\Big\}$$

where Φ is the standard normal distribution function and p_{10} and p_{01} can be written as $\sum_{j=1}^{k}(1 - \beta_{1j})\beta_{2j}$ and $\sum_{j=1^k} \beta_{1j}(1 - \beta_{2j})$, respectively. When the power of the original study is equal to the power of the replication study for all k studies, $p_{10} = p_{01}$. The power of McNemar's test can then be reduced to $\Phi\{-z_{1-\alpha/2}\}$. Assuming that $z = 1.96$, we know the power is 0.05. Thus, when the true effects of the original and replication studies are different, McNemar's test will only reject the null hypothesis 5% of the time. In other words, the false validation rate is roughly 95%.

Analysis of the t-test

Like McNemar's test, the power of the t-test to detect if the p-values of the original studies are different from those of the replication studies depends upon the power of both the original and replication studies. Consider a single study (either original or replication), with the null hypothesis $H_0 : \theta_i = 0$. The finding from this study is significant if $\frac{|T_i|}{\sqrt{v_i}} \geq 1.96$ which is equivalent to $p_i \leq 0.05$ since $p_i = \Phi^{-1}\Big(\frac{|T_i|}{\sqrt{v_i}}\Big)$ where Φ^{-1} is the standard normal quantile function.. Assuming that the variance is known for such a study and is $\frac{4}{n}$ where n is the sample size then, the power to detect if θ_i is statistically different from 0 depends on the estimate T_i and the sample size n. Extending this to the test statistic for the t-test, we note that the mean of the paired difference of p-values can be written as

$$\frac{1}{k}\sum_{j=1}^{k}\Big[\Phi^{-1}\Big(\frac{|T_{1j}|}{\sqrt{v_{1j}}}\Big) - \Phi^{-1}\Big(\frac{|T_{2j}|}{\sqrt{v_{2j}}}\Big)\Big]$$

Given that this mean takes into account both the estimate and sample size of the original and replication studies, the power to detect a difference in p-values depends on the power of both the original and replication studies.

Again, we consider the false invalidation and false validation for the t-test. A false invalidation error occurs when the effect sizes from the original and replication studies are equal but the t-test concludes that their p-values are significant different from one another. This is likely to occur when the power of the original and replication studies differ. In a simple thought experiment, we consider a single study pair. The t-test used to compare the original and replication findings in this pair would have a test statistic of $\frac{p_1-p_2}{\sqrt{V_{p_1-p_2}}}$. If we fix the parameters of the original study and let the sample size of the replication studies tend to ∞, the variance of the replication estimate will tend to 0 and, as a result, p_2 will also tend to 0. This would not be unusual in a real-world appication as researchers often increase replication sample sizes to get a more precise estimate of the original finding. In this scenario, the t-test stastic depends only on the p-value of the original study $\left(\frac{p_1}{\sqrt{V_{p_1}}}\right)$ and the error rate tends to 1. In other words, the t-test will rarely conclude that the two p-values are similar to one another. Another way to think about this error rate is to consider the case where the original and replication test statistics differ (i.e. $\frac{T_1}{\sqrt{V_1}} \neq \frac{T_2}{\sqrt{V_2}}$). If the original and replication studies have different power to detect the same effect, the numerators of these statistics will be similar, but the sample sizes and variances will differ. As a result, the p-values from the original and replication studies will be quite different, and once again the false invalidation rate will tend towards 0.80.

A false validation occurs when $\theta_{1j} \neq \theta_{2j}$ but the powers are the same: $(1-\beta_{1j}) = (1-\beta_{2j})$. The power of the original and replication studies is determined by the effect size and the sample size. When the effect size is small, the sample size needs to be relatively large in order for the test to be well-powered. If two studies have different effect sizes but the same power, their test statistics can conceivably be identical (i.e. $\left(\frac{T_1}{\sqrt{V_1}} = \frac{T_2}{\sqrt{V_2}}\right)$). For example, the study with the smaller effect size could have a larger sample size, reducing its variance while the study with the larger effect size could have a smaller sample size, resulting in a larger variance such that $cV_{small} = V_{big}$. Then, both test statistic would be similar enough to produce similar p-values, even if their effect sizes are wildly different from one another. In that case, a t-test would fail to reject the null hypothesis despite the fact that the true effect sizes are different from one another. We posit that the power of this test will tend towards 0.05 (i.e. the false validation rate tends to 0.95).

Simulation Results

In data simulations to test our analysis of the false invalidation rate, we fix the effect size of the original and replication studies at either a small, medium, or large effect. As is convention, we consider $\theta_{ij} = 0.2$ to be a small effect, $\theta_{ij} = 0.5$ to be a medium effect, and $\theta_{ij} = 0.8$ to be a large effect. For each of these effect size specifications, we then consider the case where (1) the original study is powered at 40% and the replication study is powered at 60%, (2) the original study is powered at 60% and the replication study is powered at 80%, and (3) the original study is powered at 40% and the replication study is powered at 80%. These power specifications are consistent with the range observed in the OSC data. For each of these data specifications, we run 10,000 simulations and calculate the proportion in which the t-test or McNemar's test reject the null hypothesis in favor of the alternative. The results of these simulations are displayed in Table 1. As expected, the probability of rejecting a true null hypothesis ranges from 0.8 to 1.0 for the two tests. This probability does not vary for different values of θ but does vary depending on the difference in power. The false invalidation rate and the difference in powers are not linearly related, as evidenced by the different results between the first two columns. That is, we observe different false invalidation rates in the first two columns of Table 1 despite the difference in power between original and replication studies being identical across these two columns. That being said, as the disparity in power between the original and replication studies grows larger, the false invalidation rate tends towards 1. This result is consistent with the thought experiment presented in the analysis of the t test.

	Power = 40/60	**Power = 60/80**	**Power = 40/80**
$\theta = \mathbf{0.2}$			
McNemar's	0.808 (0.01)	0.873 (0.01)	0.999 (0.0003)
T test	0.849 (0.01)	0.825 (0.01)	0.999 (0.0002)
$\theta = \mathbf{0.5}$			
McNemar's	0.808 (0.01)	0.874 (0.01)	0.999 (0.0002)
T test	0.851 (0.01)	0.826 (0.01)	0.999 (0.0002)
$\theta = \mathbf{0.8}$			
McNemar's	0.804 (0.01)	0.874 (0.01)	0.999 (0.0001)
T test	0.851 (0.01)	0.824 (0.01)	1.000 (0.0000)

Table 1: Probability of rejecting a true null hypothesis ($\theta_{1j} = \theta_{2j}$) for the t test and McNemar's test. Standard errors are reported in parentheses and rounded to the first significant digit.

In data simulations to test our analysis of power of the test, we fix the power of the original and replication studies at either 40%, 60%, or 80%. For each of these power specifications, we then consider the case where (1) the original study has a small effect size and the replication study has a medium effect size, (2) the original study has a medium effect size and the replication study has a large effect size, and (3) the original study has a small effect size and the replication study has a large effect size. For each of these data specifications, we run 10,000 simulations and calculate the proportion in which the t test or McNemar's test reject the null hypothesis in favor of the alternative. The results of these simulations are displayed in Table 2. As expected, the probability of rejecting a false null hypothesis is approximately 0.05 for both the t test and McNemar's test. That is, the false validation rate for both tests is 95%.

	$\theta = \mathbf{0.2/0.5}$	$\theta = \mathbf{0.5/0.8}$	$\theta = \mathbf{0.2/0.8}$
40% Power			
McNemar's	0.049 (0.007)	0.050 (0.007)	0.049 (0.007)
T test	0.050 (0.007)	0.051 (0.007)	0.050 (0.007)
60% Power			
McNemar's	0.050 (0.007)	0.050 (0.007)	0.049 (0.007)
T test	0.050 (0.006)	0.049(0.008)	0.049 (0.006)
80% Power			
McNemar's	0.051 (0.008)	0.051 (0.007)	0.051 (0.008)
T test	0.048 (0.008)	0.048 (0.007)	0.048 (0.007)

Table 2: Probability of rejecting a false null hypothesis for the t test and McNemar's test. Standard errors are reported in paraentheses and rounded to the first significant digit.

Appendix

Given that McNemar's test and the t test have similar properties, it is worth considering the rate at which the two tests agree on conclusions about replication. To determine whether McNemar's test and the t test always reach the same conclusions when $\theta_{1j} = \theta 2j$, we ran 1,000 trials on the same simulated data and calculated the proportion of trials in which both tests rejected their null hypothesis. The results of these simulations are found in Table 3. Similarly, to determine whether the two tests always reach the same conclusions when $\theta_{1j} \neq \theta_{2j}$, we ran 1,000 trials on the same simulated data and calucated the porpotion of trials in which both tests failed to reject their null hypothesis. The results of these simulations are found in Table 4.

	Power = 40/60	**Power = 60/80**	**Power = 40/80**
$\theta =$ **0.2**	0.980	0.966	1.000
$\theta =$ **0.5**	0.994	0.980	1.000
$\theta =$ **0.8**	0.966	0.973	1.000

Table 3: Rate of agreement between McNemar's test and the t test when $\theta_{1j} = \theta 2j$

	$\theta =$ **0.2/0.5**	$\theta =$ **0.5/0.8**	$\theta =$ **0.2/0.8**
40% Power	0.993	0.987	0.997
60% Power	0.980	0.995	0.999
80% Power	0.986	0.993	0.980

Table 4: Rate of agreement between McNemar's test and the t test when $\theta_{1j} = \theta 2j$