



# Greater Transparency Can Help Reduce Type I and Type II Errors in Research

LORNE CAMPBELL

READ REVIEWS

WRITE A REVIEW

**CORRESPONDENCE:**

[lcampb23@uwo.ca](mailto:lcampb23@uwo.ca)

**DATE RECEIVED:**

June 25, 2015

**DOI:**

10.15200/winn.143523.39755

**ARCHIVED:**

June 25, 2015

**CITATION:**

Lorne Campbell, Greater Transparency Can Help Reduce Type I and Type II Errors in Research, *The Winnower* 2:e143523.39755, 2015, DOI: 10.15200/winn.143523.39755

© Campbell This article is distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and redistribution in any medium, provided that the original author and source are credited.



First a brief review of how researchers typically decide if their research findings do or do not support their hypotheses:

Consider the simple example where a researcher has created a new way to teach math to 3<sup>d</sup> graders and wants to determine whether the new method is more effective than the standard approach to teaching math in this grade. He assigns the students to be taught math the standard way or to be taught using his new method. After a pre-determined period of time all students take the same math competency test, and the researcher conducts statistical tests to compare math scores between the two groups of students. The null hypothesis is that there is no difference in the effectiveness of the teaching methods (test scores equal across two test groups), whereas the experimental (or alternative) hypothesis is that the new teaching method is more effective than the standard method (test scores will be higher for the students taught using the new compared to the standard method). The researcher looks at the test scores in the two groups and applies a statistical test that provides the probability of getting the results in the current sample given the null hypothesis is true. The researcher then makes a decision regarding the effectiveness of his new method relative to the standard teaching method, taking into consideration information about the methods and sample, as well as the results of the statistical test(s) used. The researcher subsequently attempts to communicate this decision to the broader academic community via a manuscript submitted for publication, contingent on the evaluation of the research by a few peers and a journal editor.

In this type of research process, at least two types of errors can be made regarding the decision the researcher makes after considering all of the evidence. These errors are known as type I and type II errors:

*Type I error:* deciding to reject the null hypothesis when in fact it is correct (deciding the new teaching method is better than the standard method, when in fact *it is not better*).

*Type II error:* failing to reject a false null hypothesis, or deciding there is no effect when in fact an effect exists (deciding the new teaching method is not better than the standard method, when in fact *it is better*).

Given that a goal of science is to accumulate accurate explanations of how our world actually works, both types of errors are problematic. Finding ways to reduce these errors is therefore important for scientific discovery.

A lot of attention the past few years has focused on reducing type I errors (e.g., Simmons, Nelson &

Simonsohn, 2011, and many, many others) using both methodological (e.g., pre-registering study hypotheses, increasing sample sizes) and statistical (e.g., minimizing “p-hacking” during data analysis) approaches. Less attention has focused on how to reduce type II errors specifically. With respect to statistical tests, when the probability of correctly rejecting a false null hypothesis is low (i.e., low statistical power), the probability of making a type II error increases (if relying only on results of the statistical test to make research decisions). Increasing statistical power therefore reduces the probability of statistical tests failing to reach the chosen threshold of “statistical significance” when an effect truly exists (i.e., type II errors). There are three factors that have a big influence on the statistical power of a test:

- size of effect—smaller effects can be more challenging to detect compared to larger effects
- size of sample—smaller samples, all else being equal, provide lower power compared to larger samples
- when alpha is lower—in psychology the norm is to use an alpha level of .05. All else being equal, lower alphas decrease the probability of making a type I error but increase the probability of making a type II error compared to higher alphas

Ideally, therefore, researchers should recruit large samples of participants to increase power to help decrease type II errors in statistical tests, particularly given that the true effect sizes of interest are often unknown in advance. For example, if the researcher in the teaching method example above had 20 students in each teaching condition, the size of the effect would need to be  $d > .90$  (or rather large) in order to have 80% power to detect a difference between the two groups (see Simmons, Nelson & Simonsohn, 2013). And, again, researchers should remain mindful of the effect lowering the alpha level of their tests has on the likelihood of a type II error.

It is important to remember, however, that results of statistical tests do not dictate the decisions researchers make regarding the presence or absence of effects (see Gigerenzer & Marewski, 2015). Whatever the results of the statistical analyses used to test hypotheses, researchers need to weigh all relevant evidence, statistical as well as methodological, to reach a verdict on the perceived strength of the evidence to reject or not reject the null hypothesis and/or plan additional tests of the hypothesis. In the teaching method example, if students in the new teaching condition happened to come from schools specializing in math and science whereas students in the standard teaching condition happened to come from schools specializing in the arts (i.e., non-random assignment to condition), a significant difference in test scores in the predicted direction would not be taken as strong evidence for rejecting the null hypothesis; the lack of random assignment in this case greatly increases the risk of type I error. Similarly, a non-significant difference in test scores may not speak to the ineffectiveness of the new teaching method if the researcher was only able to recruit, for example, 20 students per teaching condition and the size of the effect turned out to be small (thus increasing the risk of type II error). In these hypothetical research scenarios, it is very easy to see how methodological limitations (when known) can influence the deliberations regarding rejecting the null hypothesis and how results of statistical tests should not alone dictate this decision making process.

The research process is, of course, not always as simple as presented in these examples. Developing hypotheses takes time and effort, as does developing ways to test hypotheses. Running studies and collecting data, as well as getting data ready for analyses (e.g., “cleaning” the data set) and conducting the analyses, also take time and effort. Importantly, researchers make many decisions during this entire process. The researcher, of course, is privy to all of these decisions given that she or he is the one making these decisions along the way. Editors and reviewers of academic journals, as well as consumers of published research, are only privy, however, to what the researcher chooses to share of the research process. Typically, what the researcher chooses to openly share of the research process occurs after all of these decisions have been made, and such sharing traditionally occurs via a manuscript submitted for peer review and ultimately publication in academic journals. In journal articles researchers tend to share the outcomes of the research process (e.g., statistically significant results in support of hypotheses) more so than the details of the research process (e.g., hypotheses developed prior to data collection and/or analyses, all study procedures and materials, pilot testing of

experimental procedures, what analyses were confirmatory or exploratory). *There is presently, therefore, not a high degree of transparency in the research process.*

Given that decisions regarding the ability of a study to reject or fail to reject the null hypothesis can only be based on information available for evaluation, having available fewer details of the research process can add more uncertainty and error to this process. For example, if hypotheses were partly based on exploratory analyses of a data set, and this information was not made publicly available, reviewers, and subsequently consumers, of the research may conclude that the results provide stronger confirmatory evidence of the hypotheses than they really do (risk of type I error). Also, consider the example of a researcher that fails to find statistical support for an innovative intervention targeted toward alleviating depressive symptoms, but does not share information, for example, regarding unequal pre-treatment depression scores across study conditions (i.e., initial depression scores happened to be lower in the standard care condition compared to the new treatment or control conditions). Reviewers, and subsequently consumers, of the research, may conclude that the results provide stronger evidence of the ineffectiveness of the new intervention than warranted (risk of type II error).

Properly evaluating scientific claims benefits from having access to more information of the entire research process (see Campbell, Loving & LeBel, 2014), information that *is* available to researchers when deciding on the strength of evidence to reject or fail to reject their own null hypotheses. This very same information, however, *is not* typically available to other researchers (and consumers of research) when making their own decisions regarding the theoretical, statistical, and practical significance of the findings reported by the researcher. Making the research process itself more transparent thus represents one important way to reduce the rates of both false positives *and* false negatives in science. Or as we say in our recent article: “Transparency in the research process, therefore, is an essential component of the scientific method because it is the only window through which we have access to this process.” (Campbell et al., 2014).

## References

- Campbell, L., Loving, T.J., & LeBel, E.P. (2014). Enhancing transparency of the research process to increase accuracy of findings: A guide for relationship researchers. *Personal Relationships*, 21, 531-545.
- Gigerenzer, G., & Marewski, J.N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41, 421-440.
- 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.
- Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2013). Life after P-Hacking. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17-19 January 2013. Available at SSRN: <http://ssrn.com/abstract=2205186> or <http://dx.doi.org/10.2139/ssrn.2205186>.