Questionable Research Practices: Definition, Detect, and Recommendations for Better Practices

Ulrich Schimmack

January 24, 2015
https://replicationindex.wordpress.com/2015/01/24/questionable-research-practices-definition-detect-and-recommendations-for-better-practices/

The term “questionable research practices” (QRPs) was popularized in an article by John, Loewenstein, and Prelec (2012).

The authors distinguish between fraud and QRPs.

Fraud is typically limited to cases in which researchers create false data.

In contrast, QRPs typically involve the exclusion of data that are inconsistent with a theoretical hypothesis. QRPs are treated differently than fraud because QRPs can sometimes be used for legitimate purposes.

For example, a data entry error may produce a large outlier that leads to a non-significant result when all data are included in the analysis. The results are significant when the outlier is removed. Statistical textbook often advise to exclude outliers for this reason. However, removal of outliers becomes a QRP when it is used selectively. That is, outliers are not removed when a result is significant or when the outlier helps to produce a significant result, but outliers are removed when removal of outliers helps to get a significant result.

The use of QRPs is damaging because published results provide false impressions about the replicability of empirical results and misleading evidence about the size of an effect.

Below is a list of QRPs.

1 **Selective reporting of (dependent) variables.** For example, a researcher may include 10 items to measure depression. Typically, the 10 items are averaged to get the best measure of depression. However, if this analysis does not produce a significant result, the researcher can conduct analyses of each individual item or average items that trend in the right direction. By creating different dependent variables after the study is completed, a researcher increases the chances of obtaining a significant result that will not replicate in a replication study with the same dependent variable.

A simple solution to preventing this QRP is to ask authors to use well-established measures as dependent variables and/or to ask for pre-registration of all measures that are relevant to the test of a theoretical hypothesis (i.e., it is not necessary to specify that the study also asked about handedness because handedness is not a measure of depression).

2 **Deciding whether to collect more data after looking to see whether the results will be significant.** It is difficult to distinguish random variation from a true effect in small samples. At the same time, it can be a costly waste of resources (or even unethical in animal research) to conduct studies with large samples, when the effect can be detected in a smaller sample. It is also difficult to know a priori how large a sample should be to obtain a significant result. It therefore seems reasonable to check data while they are being collected for significance. If an effect does not seem to be present in a reasonably large sample size, it may be better to abandon a study. None of these practices are problematic unless a researcher constantly checks for significance and stops data collection immediately after the data show a significant result. This practice capitalizes on sampling error and the experiment will typically stop when sampling error inflates the true effect size.

A simple solution to this problem is to set some a priori rules about the end of data collection. For example, a researcher may calculate sample size based on

a rough power analysis. Based on an optimistic assumption that the true effect is large, the data will be checked when the study has 80% power for a large effect (d = .8). If this does not result in a significant result, the researcher continues with the revised hypothesis that the true effect is moderate and then checks the data again when 80% power for a moderate effect is reached. If this does not result in a significant result, the researcher may give up or continue with the revised hypothesis that the true effect is small. This procedure would allow researchers to use an optimal amount of resources. Moreover, they can state there sampling strategy openly so that meta-analysts can make corrections for the small amount of biases that is still introduced by this reasonable form of optional stopping.

3 **Failing to disclose experimental conditions.** There are no justifiable reasons for the exclusion of conditions. Evidently, researchers are not going to exclude conditions that are consistent with theoretical predictions. So, the exclusion of conditions can only produce results that are overly consistent with theoretical predictions. If there are reasonable doubts about a condition (e.g., a manipulation check shows that it did not work), the condition can be included and it can be explained why the results may not conform to predictions).

A simple solution to the problem of conditions with unexpected results is that researchers may include too many conditions in their design. A 2 x 2 x 2 factorial design has 8 cells, which allows for 28 comparisons of means. What are the chances that all of these 28 comparisons produce results that are consistent with theoretical predictions?

Another simple solution is to avoid the use of statistical methods with low power. To demonstrate a three-way interaction requires a lot more data than to demonstrate that a pattern of means is consistent with an a priori theoretically predicted pattern.

4 **In a paper, reporting selectively studies that worked.**

There is no reason for excluding studies that did not work. Excluding studies that were planned as demonstrations of an effect need to be reported. Otherwise the published evidence provides an overly positive picture of the robustness of a phenomenon and effect sizes are inflated.

Just like failed conditions, failed studies can be reported if there is a plausible explanation why it failed whereas other studies worked. However, to justify this claim, it should be demonstrated that the effects in failed and successful studies are really significantly different (a significant moderator effect). If this is not the case, there is no reason to treat failed and successful studies as different from each other.

A simple solution to this problem is to conduct studies with high statistical power because the main reason for failed studies is that studies have low power. If a study has only 30% power, only one out of three studies will produce a significant result. The other two studies are likely to produce a type-II error (not show a significant result when the effect exists). Rather than throwing away the two failed studies, a researcher should have conducted one study with higher power. Another solution is to report all three studies and to test for significance only in a meta-analysis across the three studies.

5 **In a paper, rounding off a p-value just above .054 and claim that it is below .05.** This is a minor problem. It is silly to change a p-value, but it does not bias a meta-analysis of effect sizes because researchers do not change effect size information. Moreover, it would be even more silly not to change the p-value and conclude that there is no effect, which is often the case when results are not significant. After all, a p-value of .054 means that the effect in this study would have occurred if the true effect is zero or has the opposite sign.

Moreover, this problem should arise very infrequently. Even if a study is underpowered and has only 50% power, only 2% of p-values are expected to fall into the narrow range between .050 and .054.

6 **In a paper, reporting an unexpected finding as having been predicted from the start.** I am sure some statisticians disagree with me and I may be wrong about this one, but I simply do not understand how a statistical analysis of some data cares about the expectations of a researcher. Say, I analyze some data and find a significant effect in the data. How can this effect be influenced by the way I report it later? It may be a type-I error or it is not a type-I error, but my expectations have no influence on the causal processes that produced the empirical data. I think the practice of writing exploratory studies as if they were conducted an a priori hypothesis is considered questionable because it often requires other QRPs (e.g., excluding additional tests that didn't work) to produce a story that is concocted to explain unexpected results. However, if the results are presented honestly and one out of five predictor variables in a multiple-regression is significant at $p < .0001$, it is likely to be a replicable finding, even if it is presented with a post-hoc prediction.

7 **In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do).** Again, this is a relatively minor point because it only speaks about potential moderators of a reported effect. Moderation is important, but the conclusion about the main effect remains unchanged. For example, if an effect exists for men, but not for women, it is still true that on average there is an effect. Furthermore, a more common mistake is often to claim that gender or other factors did not moderate an effect based on an underpowered comparison of 10 men and 30 women in a study with 40 participants. Thus, false claims about moderating variables are annoying, but not a threat to the replicability of empirical results.

In conclusion, the most problematic research practices that undermine the replicability of published studies are selective reporting of dependent variables, conditions, or entire studies, and optional stopping when significance is reached. These practices make it possible to produce significant results when a study has insufficient power.

John et al. (2012) aptly compared these QRPs to the use of doping in sports.

Whether scientific organizations should ban QRPs just like sports organizations ban doping is an interesting question.