

Political Science Replication

Twitter: @PolSciReplicate | Facebook:

<https://www.facebook.com/PoliticalScienceReplicationBlog>

APR 19 2016

LEAVE A COMMENT

BY POLITICALSCIENCE REPLICATION.WORDPRESS.COM DATA SHARING, FAILED REPLICATIONS, JOURNALS & REPLICATION

Coding errors can be avoided

An article in the *American Journal of Political Science* was corrected after the coding of a political attitude variable was accidentally the wrong way around. Pre-publication cross-checks by the authors and the journal, as well as publication of the original data and variable transformations can avoid such problems.

The original paper with faulty codings, "Correlation not Causation: The Relationship between Personality Traits and Political Ideologies" (<http://Correlation not Causation: The Relationship between Personality Traits and Political Ideologies>)" was corrected after a new paper titled "[Personality correlates of sociopolitical attitudes in the Big Five and Eysenckian models](http://www.sciencedirect.com/science/article/pii/S019188691630232X)" (<http://www.sciencedirect.com/science/article/pii/S019188691630232X>)" mentioned inconsistencies. Even though the original work was highly cited, no one discovered the errors ("the coding errors were not identified by any other group").

Interestingly, the replicators did not frame their work as a replication, nor did they make a big fuss about it. They merely stated that they were "demonstrating that the authors had miscoded their religious and ideological measures, mistaking the conservative and religious pole for the liberal and secular pole". They made the original authors aware, moved the error discussion to an appendix, and extended their analysis based on corrected data. This procedure is commendable, because it works against fears of replication witch hunts.

The [Erratum](http://onlinelibrary.wiley.com/doi/10.1111/ajps.12216/full) (<http://onlinelibrary.wiley.com/doi/10.1111/ajps.12216/full>) to the article reads: "The interpretation of the coding of the political attitude items in the descriptive and preliminary analyses portion of the manuscript was exactly reversed. Thus, where we indicated that higher scores in Table 1 (page 40) reflect a more conservative response, they actually reflect a more liberal response. Specifically,

in the original manuscript, the descriptive analyses report that those higher in Eysenck's psychoticism are more conservative, but they are actually more liberal; and where the original manuscript reports those higher in neuroticism and social desirability are more liberal, they are, in fact, more conservative."

This error and the replication study were discussed by science writer Rolf Degen on [twitter](https://twitter.com/DegenRolf/status/718881870189084672) (<https://twitter.com/DegenRolf/status/718881870189084672>) and on [Political Science Rumors](http://www.poliscirumors.com/topic/hatemi-and-verhulst-debunked) (<http://www.poliscirumors.com/topic/hatemi-and-verhulst-debunked>). For example, one commenter asked: "Isn't this the kind of mistake that actually requires a retraction? This isn't a judgment call about model specification." Another stated, that the error could have been detected earlier on by the journal "if they had submitted the raw (de-identified) data, and the R/Stata syntax they used for recoding those variables."

Reading the articles, I had the exact same thoughts:

1. There are many corrections in political science, but not many rejections ([LaCour](https://politicalsciencereplication.wordpress.com/2015/05/30/lacour-overview-over-the-debate/) (<https://politicalsciencereplication.wordpress.com/2015/05/30/lacour-overview-over-the-debate/>) was an exception). What determines the forgiveness in our field, and is it good or bad? [I'll blog on this another time.]

2. How can we avoid simple coding errors?

Let's talk about the latter.

First, authors should conduct a replication of all the code with exactly the same data **before submitting to a journal**. For this purpose, the raw, untouched data need to be stored separately from the analysis. The syntax or Rcode to clean and transform the variables must be accessible and well commented. Ask your research assistant, co-author or someone in your lab to double check not only the analysis, but also the very beginnings of the data production. "In-house" pre-publication checks can avoid disaster (read this [story](https://scientificbsides.wordpress.com/2015/07/15/five-selfish-reasons-for-working-reproducibly/) (<https://scientificbsides.wordpress.com/2015/07/15/five-selfish-reasons-for-working-reproducibly/>) about why "It is good we spotted the problem before we submitted the paper." in cancer research). By the way, technically this is a [duplication](http://onlinelibrary.wiley.com/store/10.1111/insp.12104/asset/supinfo/insp12104-sup-0001-appendix.pdf?v=1&s=bb370bb22918a012931b039badf7a6de411d8b0a) (<http://onlinelibrary.wiley.com/store/10.1111/insp.12104/asset/supinfo/insp12104-sup-0001-appendix.pdf?v=1&s=bb370bb22918a012931b039badf7a6de411d8b0a>) (using the same data and code).

Second, **journals** should do the same cross-checks. Ideally, not only for the analysis, but also for steps from the raw data to the transformed variables that enter the models. This is a lot of work, but simply re-running code on already cleaned data will only tell you so much. The *American Journal of Political Science* actually [already conducts](https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/) (<https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/>) such pre-publication checks, but either this article was published before they started it, or they did not do the full re-analysis to detect the coding errors. So even though AJPS is a pioneer with its pre-publishing checks, more can be done.

Third, in an ideal case authors should make sure they **publish the original data** (untouched, unchanged) and code for the variable transformations. I know that this opens a can of worms. Where does "original data" start? How many raw data sets from open available sources such as the World Bank do we have to store at our webpages? Doesn't this go too far? If you cannot upload 20 data sets with each separate variable that you have collected in raw form, at least structure your data files accordingly on your computer. As I wrote [earlier](https://politicalsciencereplication.wordpress.com/2013/06/04/how-to-make-your-work-reproducible/) (<https://politicalsciencereplication.wordpress.com/2013/06/04/how-to-make-your-work-reproducible/>): All changes, e.g. variable transformations, should be made in the

Rcode. Never touch original raw data (changing .csv files manually). The advantage is that anyone else (and yourself) can start with the same, original data file from scratch. You can then share this for cross-checks by co-authors and make it available to anyone with doubts about your results later. Nothing establishes more credibility in your work than you being transparent and providing all information. Even about the raw data. Even if it is time consuming.

A note on the side: names

In this blog post I have deliberately discussed the papers, and not used the author names. You will see the names when you click on the links. This may make it a more complicated read, but I have found that being less personal, and more about the work itself, can help avoid skepticism against replicators. I will try to do this more often on this blog now.

Tagged [political science](#), [pre-publication replication](#), [replication](#), [reproducibility](#)

Blog at WordPress.com.