CA Journal of Cultural Analytics

March 5, 2018

Send us your null results

Andrew Piper[a]

[a]*McGill University*

ARTICLE INFO

Article DOI:

Journal ISSN: 2371-4549

ABSTRACT

A considerable amount of work has been produced in quantitative fields addressing what has colloquially been called the "replication crisis." By this is meant three related phenomena: 1) the low statistical power of many studies resulting in an inability to reproduce a similar effect size; 2) a bias towards selecting statistically "significant" results for publication; and 3) a tendency to not make data and code available for others to use.

A considerable amount of work has been produced in quantitative fields addressing what has colloquially been called the "replication crisis."[1] By this is meant three related phenomena: 1) the low statistical power of many studies resulting in an inability to reproduce a similar effect size; 2) a bias towards selecting statistically "significant" results for publication; and 3) a tendency to not make data and code available for others to use.

What this means in more straightforward language is that researchers (and the public) overwhelmingly focus on "positive" results; they tend to over-estimate how strong their results are (how large a difference some variable or combination of variables makes); and they bury a considerable amount of decisions/judgments in their research process that have an impact on the outcomes. The graph in Figure 1 down below represents the first two dimensions of this problem in very succinct form (see Simmons et al for a discussion of the third).[2]

Why does this matter for Cultural Analytics? After all, much of the work in CA is insulated from problem #1 (low power) because of the often large sample sizes used. Even small effects are mostly going to be reproducible with large enough samples. Many will also rightly point out that a focus on significance testing is not always at the heart of interpretive research. Regardless of the number of texts used, researchers often take a more descriptive or exploratory approach to their documents, where the idea of "null" models makes less sense. And problem #3 is dealt with through a code and data repository that accompanies most articles (at least in CA and at least in most cases).

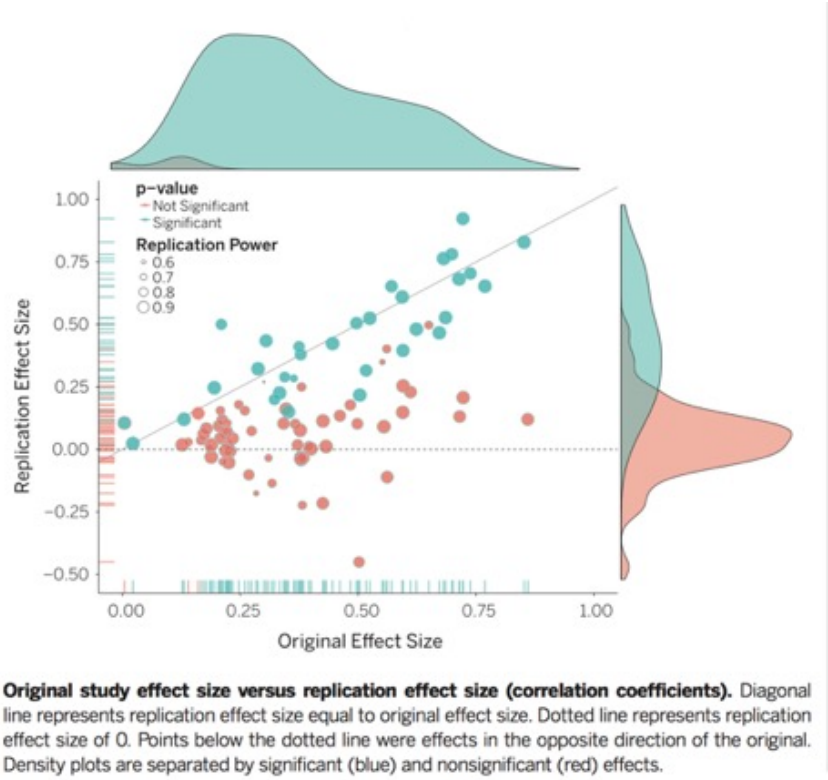
But these caveats overlook a larger and more systemic problem that has to do with selection bias towards positive results. Whether you are doing significance testing or just saying you have found something "interesting," the emphasis in publication is almost always on finding something "positive." This is as much a part of the culture of academic publishing as it is the current moment in the shift towards data-driven approaches for studying culture. There is enormous pressure in the field to report something positive -- that a method "worked" or "shows" something. One of the enduring critiques of new computational methods is that they "don't show us anything we didn't already know." While many would disagree (rightly pointing to positive examples of new knowledge) or see this as a classic case of "hindsight bias" (our colleagues' ability to magically always be right), *it is actually true that in most cases these methods don't show us anything at all*. It's just that you don't hear about those cases.

If we were to take the set of all experiments ever conducted with a computer on some texts, I would expect that in (at least) 95% of those cases the procedure yielded no insight of interest. In other words, positive results would be very rare. And yet, miraculously, all articles in CA report a positive result (mine included). To be fair, this is true of literally all literary and cultural studies. No one to my knowledge has ever published an article that said, I read a lot of books or watched a lot of television shows and it turns out my ideas about them weren't significant. But this too happens all the time. We just never hear about it.

It's time to change that culture. Researchers in other fields have made a variety of suggestions to address this issue, including pre-submitting articles prior to completion so acceptance isn't biased towards positive results, to making the research process as open and transparent as possible.[3] At CA, we want to start by encouraging submission of pieces that don't show positive results, however broadly defined. This can be another way that the journal CA, but also work in cultural analytics more broadly, can begin to change research culture in the humanities and cultural studies. It means not only changing the scale of our evidence considered or making our judgments more transparent and testable. It also means being more transparent about all the cases where our efforts yield no discernible effect or insight. As others have called for, it is time to embrace failure as an epistemic good.[4] This may be CA's most radical gesture yet in changing the culture of research in the field of cultural studies.

So let me open the floodgates here: we pledge to publish your null result. By null result, I mean either something that shows no statistical significance (i.e. using machine learning, prizewinning novels cannot be distinguished from novels reviewed in the New York Times with a level of accuracy that exceeds random guessing). Or something that shows no discernibly interesting pattern from an interpretive point of view (we ran a topic modeling algorithm on all of ECCO and regardless of the parameters used the topics do not seem to represents reasonable categories of historical interest, i.e. it didn't work very well no matter what we did).

These are examples of the kind of null results we're thinking of. I'm sure you can think of many, many more. It is important that the submission be as framed, justified and fleshed out as that positive result you've been salivating about publishing in the highest prestige place you can imagine. But just because the piece shows "nothing" (you know what I mean, don't get all postmodern on me), doesn't mean it shouldn't be published. If the question matters, then we ought to hear about how a method failed to address that question. This will not only save researchers time in knowing what to focus on, it can also open-up shared areas of inquiry—maybe there was a problem in the method that could be improved or maybe whatever you're looking for really doesn't have much of an effect. Only with repeated attempts can we ever get any confidence about spurious ideas or methodological limitations. Only then are we going to inhabit a research culture where everyone isn't always right.



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Fig. 1 The distribution on the top of the graph represents published results -- overwhelmingly biased towards statistical significance (in blue, see the little dark blue part which buries the pink insignificant studies). The distribution on the right represents replicated results, which show a normal distribution that overwhelmingly favors insignificant results (pink). As commentators have increasingly pointed out, current models for statistical inference are mathematically biased towards over-estimating effects of real-world associations. From: Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science* 349, aac4716 (2015). DOI: 10.1126/science.aac4716.

Notes

[1]A few selected readings: Barbara Spellman, "A Short (Personal) Future History of Revolution 2.0," *Perspectives on Psychological Science* 10.6 (2015): http://journals.sagepub.com/doi/full/10.1177/1745691615609918. Brian D. Earp and David Trafimow, "Replication, Falsification, and the Crisis of Confidence in Social Psychology," *Frontiers in Psychology* May 19, 2015: Â https://doi.org/10.3389/fpsyg.2015.00621 and the "Reproducibility Project" of the Open Science Foundation: https://osf.io/ezcuj/wiki/home/. For popular accounts, see Ed Yong, "Psychology's Replication Crisis Can't Be Wished Away," *The Atlantic* March 4, 2016: https://www.theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/ and Christie Aaschwanden, "Failure is Moving Science Forward," *Five-Thirty-Eight* March 24, 2016: https://fivethirtyeight.com/features/failure-is-moving-science-forward/. It is important to point out that while many of the examples relate to psychology, other fields including epidemiology and biomedicine have raised serious concerns as well.

[2]Joseph P. Simmons, Leif D. Nelson, Uri Simonsohn, "False Positive Psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science* 22.11 (2011): http://journals.sagepub.com/doi/full/10.1177/0956797611417632.

[3]See the Center for Open Science: https://cos.io, and the Open Science Framework: https://osf.io and D. Stephen Lindsay, "Replication in Psychological Science," 26.12 (2015): http://journals.sagepub.com/doi/abs/10.1177/0956797615616374.

[4]John Unsworth, "The Importance of Failure," *Journal of Electronic Publishing* (1997): http://hdl.handle.net/2142/192 and more recently, the symposium organized by Geoffrey Rockwell, "On the Benefits of Failure": https://www.digitalscholarsua.com/conference-2018/.