CA Journal of Cultural Analytics

April 1, 2019

Do we know what we are doing?

Andrew Piper[a]

[a] *McGill University*

ARTICLE INFO

Article DOI: 10.22148/001c.11826

Journal ISSN: 2371-4549

ABSTRACT

In November 2012, the newly created Open Science Collaboration published a brief article announcing a multi-year effort to "estimate the reproducibility of psychological science." The collaboration was directed by Brian Nosek of the University of Virginia and would eventually involve over 250 co-authors. According to the collaboration, reproducibility was one of, if not the single most defining feature of the social endeavor known as "science." "Other types of belief," the authors write, "depend on the authority and motivations of the source; beliefs in science do not." The ability to reproduce scientific results across time and space -- the ability to have results be *independent* of the individuals involved -- is what the authors argued makes science science. And yet the eventual findings of the reproducibility project showed a remarkable reproductive failure. Over half of all studies failed to indicate similar effects upon replication. The very value upon which science was supposed to be founded appeared to be an exception rather than a norm.

In November 2012, the newly created Open Science Collaboration published a brief article announcing a multi-year effort to "estimate the reproducibility of psychological science."[1] The collaboration was directed by Brian Nosek of the University of Virginia and would eventually involve over 250 co-authors. According to the collaboration, reproducibility was one of, if not the single most defining feature of the social endeavor known as "science." "Other types of belief," the authors write, "depend on the authority and motivations of the source; beliefs in science do not."[2] The ability to reproduce scientific results across time and space -- the ability to have results be *independent* of the individuals involved -- is what the authors argued makes science science. And yet the eventual findings of the reproducibility project showed a remarkable reproductive failure.[3] Over half of all studies failed to indicate similar effects upon replication.[4] The very value upon which science was supposed to be founded appeared to be an exception rather than a norm.[5]

Nan Z. Da's study published in *Critical Inquiry* is part of a growing body of work that seeks to introduce the idea of replication into the humanities.[6] While the practice of reproducing prior work is far from normalized in the humanities, it is to be welcomed. As editor of this journal, I am committed to fostering an environment where we can work together to assess the reliability and fruitfulness of our work. The aim of such endeavours is to arrive at a greater degree of consensual forms of knowledge about behavior in the world. As the authors of the OSC write, the value of replication, when done well, is that it can "increase certainty when findings are reproduced and promote innovation when they are not." The value of replicability is itself one of the affordances of the very methods Da seeks to critique.

And yet despite Da's aims of testing the reproducibility of computational literary research, her work fails to follow any of the procedures and practices established by replication projects like the OSC. Indeed, her article cites no relevant literature on the issue, suggesting little familiarity with the topic. While invoking the epistemological framework of replication—that is, to prove or disprove the validity of both individual articles as well as an entire field—her practices follow instead the time-honoured traditions of selective reading from the field of literary criticism. Da's work selectively collects a "handful" of articles for review; selectively cites certain aspects or figures from these articles; and selectively frames all computational methods within a single narrow definition of significance testing, which she herself does not follow. And yet from this selective stew, Da makes the broadest, most sweeping claims imaginable: computational literary studies is an invalid field.

Compare this with the conclusions of the OSC, who, after collaborating for two years with original authors and establishing rigorous standards for replication, found that over half of all experiments failed to replicate and yet still write the following, which is worth quoting in full:

> After this intensive effort to reproduce a sample of published psychological findings, how many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice. Humans desire certainty, and science infrequently provides it. As much as we might wish it to be otherwise, a single study almost never provides definitive resolution for or against an effect and its explanation. The original studies examined here offered tentative evidence; the replications we conducted offered additional, confirmatory evidence. In

> some cases, the replications increase confidence in the reliability of the original results; in other cases, the replications suggest that more investigation is needed to establish the validity of the original findings. Scientific progress is a cumulative process of uncertainty reduction that can only succeed if science itself remains the greatest skeptic of its explanatory claims.[7]

The OSC, in other words, made a good faith effort to work with other members of their field to test the reproducibility of past research; attempted to gather a reasonable cross-section of studies representative of the field; established criteria for success and failure when it came to replication; and even after finding a lower than 50% replication rate made no sweeping claims about the validity of their own field or even the individual studies under review.

Da does none of this. It is precisely this combination of highly selective evidence, inconsistent computational methods, alongside the most sweeping claims imaginable, that casts serious doubt on the credibility of Da's scholarship and raises important questions as to the qualifications of the editorial board of *Critical Inquiry* to assess such efforts as we move forward.[8]

Da's work is valuable, then, not because of the computational case it makes (that work remains to be done), but in the way it highlights with remarkable consistency a much larger problem facing the field: how are we going to combat the problem of selective reading? Or to put it in reverse, how can we begin to address the problem of *generalization*, the ways in which we move from individual observations (no matter how few or how many) to more general empirical claims about things in the world? Whether we are replicating the work of others or making novel arguments that have not been aired before, the problem of generalizability is central to scholarly knowledge production. And yet it has remained a wholly undertheorized and underdiscussed concept within the humanities.

In what follows I discuss the conceptual and methodological shortcomings of Da's work in order to illustrate the challenges that traditional critical methods face when it comes to the practice of generalization. Da's work is valuable in so far as it foregrounds so many of the problems that accompany traditional critical models of evidence that are used to make large-scale evidentiary claims and why it is this, far more than the bogeyman of computation Da invokes, that needs to be addressed as we move forward.

Samples, or Good Old-Fashioned Selection Bias

When undertaking their replication project, the OSC generated a sample of 100 studies taken from three separate journals within a single year of publication. Their rationale was the following:

> These were selected a priori in order to (i) provide a tractable sampling frame that would not plausibly bias reproducibility estimates, (ii) enable comparisons across journal types and subdisciplines, (iii) fit with the range of expertise available in the initial collaborative team, (iv) be recent enough to obtain original materials, (v) be old enough to obtain meaningful indicators of citation impact, and (vi) represent psychology subdisciplines that have a high frequency of studies that are feasible to conduct at relatively low cost.

Notice how they do not claim that their sample is a perfect representation of the entire field. No sample can be. However, they do provide a rationale for the articles included as well as all of the limitations that are associated with these choices (these studies are not the most recent, they need to match available expertise of replicators, and they are "relatively low cost" studies).

Da on the other hand chooses 14 articles from different years and different journals. Her stated rationale is the following:

> I discuss a handful of CLS arguments (chosen for their prominent placement, for their representativeness, and for the willingness of authors to share data scripts or at least parts of them).

No one would expect Da to replicate as many articles as the OSC. But notice how there is no actual concrete number of publications indicated (I arrive at 14 because there are 14 pieces from which she draws a statistic) nor is there a consistent framework for selection. Instead, Da chooses a "handful" based on what she identifies as "prominence", "representativeness", and the "willingness of authors to share data." The first and second concepts are not defined. Is one article in the

Journal of Cultural Analytics more "prominent" than another, justifying why some but not all articles from CA were included? Does prominent mean "more cited," "more downloaded," written by a person at a more prestigious institution or with more Twitter followers? All of these could be possible ways of measuring prominence but Da doesn't bother. In terms of representativeness, what are these articles representative of? Da never says. Representative of people's age? Training? Gender? Participation in a research grant? Who knows. The third point is indeed a good one, as you cannot replicate something without something to replicate. Here too though there are a number of articles in CA that do have data and code that Da chooses not to replicate. Why not? While the OSC authors indicate at great length why articles were or were not chosen, Da never specifies any of this. It's part of the black box of criticism.

Indeed, the only clear linkage appears to be that these studies all "fail" by her criteria. Imagine if the OSC had found that 100% of articles sampled failed to replicate. Would we find their results credible? Da by contrast is surprisingly only ever right. This is the evidentiary mode of the literary critic, who only picks examples to support her argument and suppresses those that don't. It's a rhetorical exercise of persuasion, not an empirical one of proof. And yet Da's work aims to make empirical claims about the validity of other observations in the world.

Da's work thus nicely highlights one of the primary failings of traditional criticism with respect to evidence—it only ever works with positive examples. When was the last time you read an article in literary or film studies that said, I watched a bunch of movies and only some of them showed this effect I'm talking about. In fact, most don't, but I think they're important anyway. A more credible practice of generalization requires that we examine both positive and negative examples in order to understand the representativeness (or distinctiveness) of the sample we are observing and the representativeness (or distinctiveness) of the individual observations within those samples. Anything less is uncredible.

Passages, or what is one CV evidence of?

We can move one layer down to questions of representation within articles as well. What portion of an article is representative of that article? If a sample is intended as a representation of a larger population or world, we need some way of deciding what aspects of an article (or any document) are representative of that article (or document). This too is an extremely challenging task. Once again, the OSC makes this process explicit when they write:

> By default, the last experiment reported in each article was the subject of replication. This decision established an objective standard for study selection within an article and was based on the intuition that the first study in a multiple-study article (the obvious alternative selection strategy) was more frequently a preliminary demonstration.

And once again the authors then spend time reflecting on the limitations of their selection criteria. For sure, failing to replicate the final experiment of an article does not discredit the whole study. The OSC authors acknowledge as much, which is why their ultimate benchmark is experiments ("studies" in their words) not articles.

Not so Da. She never specifies what exactly she will be replicating in each article, nor does she explicate under what conditions she could arrive at a replicate/failed-to-replicate judgment, nor does she specify at what point enough articles failing is satisfactory to conclude that a whole field has failed. It all happens once again inside the black box of critical judgment. Not unsurprisingly she proceeds to apply different criteria to every article, making debatable methodological choices along the way, as well as numerous errors, that are clearly designed to foreground differences.[9] She misnames authors of articles, mis-cites editions, mis-attributes arguments to the wrong book, and fails at basic math.[10] And yet each of these assertions always adds-up to the same certain conclusion: failed to replicate.

Let me provide some examples. For example, in my own work she highlights the fact that just because Augustine's *Confessions* exhibits a certain lexical pattern does not mean that all texts that do so are "conversional" in my terms.[11] Precisely. Which is why I spend over 6,000 words in the article validating whether the novels so identified can be labeled as such and provide a figure of the importance of doing so (see Fig. 2 in the original article). She then makes much of stemming the Latin text and normalizing the PCA for visualization and proceeds to produce a graph that largely reproduces the original findings in my article, which hinges on the strong partition between the pre- and post-conversional books. (The fact that she claims that the later books aren't related to Augustine's conversion because they "are about Genesis" would give Augustine scholars a good chuckle.) It is notable that much of Da's analysis of articles throughout her piece conditions on PCA, which is one small piece of the text-analytical toolkit. Indeed, in my own piece it is only used for purposes of visualization and not for the actual calculations used in the final model I construct to measure conversionality in novels -- which Da never addresses.

In her critique of my piece on Wertherness in Goethe's corpus co-authored with Mark Algee-Hewitt, she claims that we fail to provide a null model to prove causality (i.e. that *Werther* actually influenced these later works).[12] That would indeed be a challenge. But we set out to do no such thing. As Mark Algee-Hewitt writes in his response to Da, "Even a cursory reading of the article reveals that we are not interested in questions of the 'influence of *Werther* on other texts': rather we are interested in exploring the effect on the corpus when it is reorganized around the language of *Werther*. The topology creates new adjacencies, prompting new readings: it does not prove or disprove, it is not right or wrong - to suggest otherwise is to make a category error."[13] Our work is inspired by generations of literary theory on interextuality that explores literary writing for associations and relations between language. This approach is entirely fitting within the normal tradition of hermeneutic close reading, only using new forms of textual mediation. And it is entirely fitting within the paradigm of "exploratory data analysis" in the field of statistics. In other words, Da misreads as both a humanist and a statistician.

Perhaps even more telling, is the way Da only chooses to discuss articles of mine where I explicitly don't do hypothesis testing and ignores all of the other ones where I do in order to generalize about the absence of hypothesis testing in either my work or the field at large, that is, to fit the data to her hypothesis.[14] This is the kind of selective data analysis that is deeply suspicious. Worse still, Da's accusations about the failure of (some) articles in CLS to do hypothesis testing are made in an article that itself lacks any such testing. There appears to be one standard for Da and one for everyone else.

Let me use one final example of the strategic value of Da's selectivity for her work. Da makes a claim that money is being wasted on CLS. Given that software is free, she says, it "begs the question of why we need 'labs' or the exorbitant funding that CLS has garnered." Her evidence for this exorbitant funding is a link to my CV in footnote 5. In what context is "one CLS author" sufficient evidence of funding *for a field*?

But perhaps Da's insinuation is that I am personally being wasteful. Here too we could use evidence instead of insinuation. Had she inquired with the Social Science and Humanities Research Council of Canada, she would have learned that 50.5% of my personal funding last year was dedicated to funding students. At private U.S. institutions like Da's (or *Critical Inquiry*'s home institution), graduate students in the humanities are most often funded independently from departments and faculty, subsidized through the high cost of undergraduate tuition or gifts by wealthy donors that are used to maintain access to elite higher education for their descendants. In the public system of Canada, our students in both the arts and sciences are largely funded by the individual research grants of faculty. I am very proud of the fact that with one grant alone I have financially supported over 75 students in the past five years. Much of the rest of the funding is either transferred to partner institutions, used on administrative costs, or used to organize annual conferences (is Da opposed to conference travel I wonder, because this is definitely worth debating). But Da's goal is to smear not understand. She exhibits a basic lack of knowledge about educational funding in Canada and the nature of large-scale research funding in general. Nor does she disclose where her own funding came from to pay for her research assistants. For sure, there is a robust debate to be had about the best use of academic funding. But a link to a CV in a footnote is no way to have it. This is reminiscent of conservative tactics of citing MLA paper titles out of context: "Intransitive Encounter"? what a bunch of mumbo jumbo. It has no place in academic discourse.

Methodological Plurality, or the Perspectival Nature of Knowledge

Perhaps the greatest limitation of Da's piece is the extremely narrow definition of statistical inference and computational modeling she applies to the pieces under review. In Da's view, the only appropriate way to use data is to perform what is known as significance testing, where we use a statistical model to test whether a given hypothesis is "true."[15] There is no room for exploratory data analysis, for theory building, or predictive modeling (machine learning) in her view of the field. This is particularly ironic given the fact that Da herself performs no such tests! She holds others to standards to which she herself is not accountable. Nor is it an accurate account of the state of the field, where numerous articles have been written about the limits of precisely the methods Da seems to prize (though not practice).[16]

For Da, there is only a single way to read. This way is at once homogenous and totally unspecified. There is no mediation, no situatedness, and for sure no diversity to Da's model of reading that she puts forward in her piece as an antidote to computational reading. As numerous articles on the importance of computational modeling have argued, textual knowledge is always situated and dependent on the observer's position in the world and the tools, techniques, and technologies through which their point of view is constructed.[17] Modeling makes that positionality explicit. Da's own position, the assumptions guiding her choices and interpretations, is left entirely unexplicated in her piece.

In its narrowness, Da's work appears deeply out of touch with existing research related to reading, meaning, and probabilistic modeling from other fields. For example, she makes much of what she sees as a central limitation of computational methods: the recourse to counting words ("CLS papers make arguments based on the number of times x word or gram appears"; " In CLS data work there are decisions made about which words or punctuations to count and decisions made about how to represent those counts. That is all."). And yet decades of work in computational linguistics has drawn attention to the analytical value of modeling language use probabilistically as a way of modeling human behavior and judgment.[18] This is by no means a settled matter, but to suggest that probabilistic models have *no* bearing in assessing textual meaning is obtuse. Da wants us to believe that even if these models work for other kinds of documents, they don't work on literary texts. Where is the evidence that probabilistic models work for every other document in the world, except the ones that are fictional in their content and that Da conveniently wishes to be the exclusive expert on? Where scholars are trying to infer meaning from distributions of word probabilities (never individual words), it is important to emphasize that not only are such representations potentially good approximations of human judgments about texts and meaning, but that such reductive representations of texts or ideas are the necessary cost of scale. You cannot scale up your claims without paying a price. See Bruno Latour.

Nan Da says that data is inappropriate for the study of literature. No if's, and's or but's. She says we would be better off just reading more books. The point I want to underscore (and have been making for some time) is that Da's solution of *only* reading books by hand is of limited value for making *certain* statements about literature. Notice how my position has two caveats while Da's has none. There is no scenario in her model of literary study that allows for quantitative evidence (or any kind of technological mediation). In my scenario, quantitative evidence can be a valuable *complement* to assist scholars in the process of generalization and help make our evidentiary claims more credible. As Da herself argues, "Typical applications of textual data mining involve a trade-off: speed for accuracy, coverage for nuance." Precisely. In order to scale-up our evidence, we have to sacrifice nuance. That is the cost of scale. If you want to make claims about "the nineteenth-century novel" or "computational literary studies," it defies credibility to do this with a few examples, no matter how complex your reading. We do indeed need to do a better job of knowing what we are doing, just not in the way Da suggests.

Conclusion

All of the limitations of Da's endeavor should not blind us to remaining skeptical of our own research. To be doubtful is a good thing. Replication and questioning of previous results belongs at the heart of scholarship. As the authors of the OSC paper write, scholarship "can only succeed if it itself remains the greatest skeptic of its explanatory claims." For sure, there have to be errors and problems with previous computational research. It would be statistically impossible for it all to be 100% reliable. This is the very definition of research as a form of "Forschung," a searching for knowledge, understanding, and insight. Research necessarily involves error and failure.

What we need then, and what I am committed to fostering with this journal, are supportive and collaborative engagements with each other's work and the work of other fields to build consensus about what we (think we) know and what we remain uncertain about and the best way to understand those things better. This requires questioning the findings of computationally driven research, but it also means testing the claims of non-computationalists with larger samples and more transparent methods. It has to work both ways.

To conclude, let me summarize what I see as some of the primary goals that data-driven literary study is trying to achieve in order to address Da's central challenge. This is my own personal list and I hope others might add more. So why am I using computation to study literature? In order to:

1. address the problem of generalization and insufficient or poorly-sampled evidence;
2. address intrinsic qualities of texts, i.e. that language use is highly repetitive and can be modeled by probability distributions;
3. identify longer-term historical patterns than traditional methods and periodizations allow for;
4. draw attention to the technological mediations that govern reading (esp. through the practice of self-reflexive modeling).

Nan Da's article is an attempt to polarize and prohibit. It is to be frank, very much of our time. I would take the exact opposite approach: bridge and build consensus. And there two roads diverged as they say.

Notes

1. Open Science Collaboration, "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science," *Perspectives on Psychological Science* 7, no. 6 (November 2012): 657-60. DOI:10.1177/1745691612462588. Open Science Collaboration, "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science" 657.
2. Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science* 28 Aug 2015: Vol. 349, Issue 6251, aac4716. DOI: 10.1126/science.aac4716.
3. According to the authors of the report, "Ninety-seven percent of original studies had significant results ($p < 0.05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size."
4. For an overview of what has come to be known as the reproducibility crisis more generally, see Barbara Spellman, "A Short (Personal) Future History of Revolution 2.0," *Perspectives on Psychological Science* 10.6 (2015): doi.org/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: doi.org/10.3389/fpsyg.2015.00621. 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.
5. Nan Z. Da, "The Computational Case Against Computational Literary Studies," *Critical Inquiry*45 (Spring 2019) 601-639.
6. Open Science Collaboration, "Estimating the Reproducibility of Psychological Science" aac4716-7.
7. Da makes much in her appendix on insisting that articles in computational literary studies be reviewed by a statistician. This is insufficient -- lack of domain knowledge does not make a person trained in quantitative methods a qualified reviewer. See Da's own piece as evidence of this problem. For sure, we want reviewers qualified in computational literary studies to review pieces in computational literary studies. But Da oversimplifies what qualified means here, insinuating without evidence that such reviewers have not been actively reviewing pieces.
8. For a list, see Ben Schmidt, "A computational critique of a computational critique of a computational critique," http://benschmidt.org/post/critical_inquiry/2019-03-18-nan-da-critical-inquiry/.
9. She cites Mark Algee-Hewitt as Mark Hewitt, cites G. Casella as the author of *Introduction to Statistical Learning* when it was Gareth James, cites me and Andrew Goldstone as co-authors in the Appendix when we were not, claims that "the most famous example of CLS forensic stylometry" was Hugh Craig and Arthur F. Kinney's book that advances a theory of Marlowe's authorship of Shakespeare's plays which they in fact do not, and miscalculates the number of people it would take to read fifteen thousand novels in a year. The answer is 1250 not 1000 as she asserts. This statistic is also totally meaningless.
10. Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," *New Literary History* 46.1 (Winter 2015): 63-98.

11. Andrew Piper and Mark Algee-Hewitt, "The Werther Effect I," *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ed, Matt Erlin and Lynn Tatlock. (Rochester: Camden House, 2014) 155-184.
12. See Mark Algee-Hewitt's response in the *Critical Inquiry* forum.
13. See Andrew Piper and Eva Portelance, "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading," *Post-45*(2016); Eve Kraicer and Andrew Piper, "Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction," *Journal of Cultural Analytics*, January 30, 2019. DOI: 10.31235/osf.io/4kwrg; and Andrew Piper, "Fictionality," *Journal of Cultural Analytics*, Dec. 20, 2016. DOI: 10.22148/16.011
14. Statements like the following also suggest that she is far from a credible guide to even this aspect of statistics: "After all, statistics automatically assumes that 95 percent of the time there is no difference and that only 5 percent of the time there is a difference. That is what it means to look for p-value less than 0.05." This is not what it means to look for a p-value less than 0.05. A p-value is the estimated probability of getting our observed data assuming our null hypothesis is true. The smaller the p-value, the more unlikely it is to observe what we did assuming our initial hypothesis is true. The aforementioned 5% threshold says nothing about how often there will be a "difference" (in other words, how often the null hypothesis is false). Instead, it says: "if our data leads us to conclude that there is a difference, we estimate that we will be mistaken 5% of the time." Nor does "statistics" "automatically" assume that .05 is the appropriate cut-off. It depends on the domain, the question and the aims of modeling. These are gross over-simplifications.
15. The literature debating the values of significance testing is vast. See Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22, no. 11 (November 2011): 1359-66. doi:10.1177/0956797611417632.
16. Andrew Piper, "Think Small: On Literary Modeling." *PMLA*132.3 (2017): 651-658; Richard Jean So, "All Models Are Wrong," *PMLA*132.3 (2017); Ted Underwood, "Algorithmic Modeling: Or, Modeling Data We Do Not Yet Understand," *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*, eds. J. Flanders and F. Jannidis (New York: Routledge, 2018).