How Are We Doing? Data Access and Replication in Political Science

Ellen M. Key, *Appalachian State University*

ABSTRACT Data access and research transparency (DA-RT) is a growing concern for the discipline. Technological advances have greatly reduced the cost of sharing data, enabling full replication archives consisting of data and code to be shared on individual websites, as well as journal archives and institutional data repositories. But how do we ensure that scholars take advantage of these resources to share their replication archives? Moreover, are the costs of research transparency borne by individuals or by journals? This article assesses the impact of journal replication policies on data availability and finds that articles published in journals with mandatory provision policies are 24 times more likely to have replication materials available than articles those with no requirements.

The controversy surrounding LaCour and Green (2014) highlights the importance of replication and verification. The inability of researchers to replicate the central findings (Broockman, Kalla, and Aronow 2015) and the subsequent retraction of the article by *Science* editors caused a scandal in the field and beyond—similar to the aftermath of the discovery of Reinhart and Rogoff's (2010) spreadsheet error in economics (Herndon, Ash, and Pollin 2013). These alleged errors, and others like them, were identified using publicly available replication archives.[1] The public availability of these archives, however, is largely due to efforts made by journals to increase research transparency.

Data access and research transparency (DA-RT) is a growing concern for the discipline. Technological advances have greatly reduced the cost of sharing data, enabling full replication archives consisting of data and code to be shared on individual websites, as well as journal archives and institutional data repositories. But how do we ensure that scholars take advantage of these resources to share their replication archives? Moreover, are the costs of research transparency being borne by individuals or by journals? Expanding on the work of Gherghina and Katsanidou (2013), I move from the journal-level to the article-level to assess the impact of journal replication policies on data availability. I conclude with suggestions for increasing research transparency.

EXISTING EFFORTS

The goal of publishing replication archives is not simply internal verification or the correction of sloppy scholarship. Rather, replication also allows for extension through the collection of new data and the application of different methods (Fowler 1995; King 2006). Although scholars have an incentive to ensure that their data are available and up to date as a way to increase exposure and citation counts (Gleditsch, Metelits, and Strand 2003), it is difficult to achieve compliance on a voluntary basis (Anderson et al. 2005; King 1995). Recognizing that relying on scholars to self-police is suboptimal, journals have recently created or revised their replication policies to advance social rather than individual responsibility. In other words, the burden is shifting to editors to ensure the availability of replication archives for work published in their journal (Gherghina and Katsanidou 2013; Ishiyama 2014).

Part of this shift is due to journals committing to the DA-RT statement developed by the APSA council (APSA 2014). Based on the "principle that sharing data and information fuels a culture of openness that promotes effective knowledge transfer" (Lupia and Elman 2014, 20), editors of DA-RT journals require data to be uploaded to a journal repository at the time of publication. There are many benefits of these repositories, including durable, central archives that do not require individuals to be responsible for maintenance. Older, more prestigious, and general-interest journals are more likely to have replication policies than those with lower-impact factors or more specific audiences (Gherghina and Katsanidou 2013). This is a self-reinforcing process because more readily available data increases citation counts, thereby boosting the impact factor of a journal.

Journal policies that require replication may affect material availability beyond an author's natural tendency to publish replication archives. Ensuring that replication standards are met, however, strains scarce journal resources. If scholars are already maintaining complete replication archives on their own, there is no need for editors to police their authors. If replication policies are not fully enforced, the effort expended for partial enforcement may be wasted (Dafoe 2014).

Ellen M. Key *is assistant professor of political science at Appalachian State University. She can be reached at keyem@appstate.edu.*

  doi:10.1017/S1049096516000184

RESEARCH DESIGN

To determine the ability of journal policies to affect the availability of data and replication archives, articles were examined for data and code availability, as well as for the location of replication materials. The sample consists of every quantitative[2] article from 2013 and 2014 in six leading journals: *American Political Science Review* (*APSR*), *American Journal of Political Science* (*AJPS*), *British Journal of Political Science* (*BJPS*), *International Organization* (*IO*), *Journal of Politics* (*JOP*), and *Political Analysis* (*PA*). In addition to impact factor, the journals were chosen based on scope: four of general interest, one focusing on a broad subfield, and one highly specialized subdiscipline.

Replication policies were determined based on online policies or e-mail correspondence with a journal editor if a posted policy could not be located. Although higher-impact journals are more likely to have replication policies in place, there is variation in terms of policy type: some are focused on verification and others only on data availability. *IO* has the most stringent replication policy of the six journals examined, requiring authors to provide editors with data and code for replication before publication. *AJPS* and *PA* rank second-most stringent by requiring data citation and replication materials to be uploaded to the journal's dataverse.[3] Both *BJPS* and *JOP* require an author to note the location of replication materials in an article but have no policies in place that mandate replication files to be provided. Last, *APSR*'s policy is that replication materials are to be provided by an author, but there are no requirements regarding the location of or directions to the replication archive.

It is important to note that several journals have changed their replication policies in a move toward increased transparency since the data were collected. *AJPS* now verifies analyses. *JOP* now requires replication materials to be uploaded to the journal's dataverse before publication. A change to *APSR*'s policy is forthcoming but has not yet been implemented.

Data collection for this analysis took place from October 2014 through January 2015 and focused on following the directions to replication materials found in an article, widening the search only when necessary. Due to the push for social responsibility in data storage, I first determined whether the data and code were available on a website maintained by a journal, including journal-specific dataverses and supplementary materials pages. An article was coded as being available from the journal if the data and/or code could be downloaded from the journal's website or repository. If an article indicated that the replication materials were available on an author's dataverse or other personal website, it was coded as such if the links were still functional. Materials were coded as being unavailable on the listed website if a link provided in an article directed readers to the homepage rather than to a data-access page. In the event of broken links or no mention of replication materials, a web search attempted to find an active website for an author(s). Ultimately, some web presence was found for at least one author of the remaining 242 articles.

RESULTS

Of the 586 articles published in these six journals during the period studied, 494 contained some type of quantitative data. As shown in table 1, a full replication archive—that is, data and replication code—was available for 58% (287) of the articles. The availability of replication materials varied widely by journal, from a high of 98.1% for *PA* to a low of only 32.4% for *APSR*. This variability is likely due to variation in replication policies; those journals with availability rates of more than 90% are also those that require authors to provide a data citation and upload materials to the journal's dataverse. Policies requiring mandatory provision, however, are not sufficient to ensure complete compliance. Beginning in 2014, *IO* required authors to submit data for editorial replication before publication; there was no significant change in the availability rate (i.e., approximately 90%) after the policy shift.

Of those articles that provided replication materials, a majority of authors provided both data and code. Overall, 292 (59.1%) had full replication archives and 167 (33.8%) provided neither data nor code. At 94.4%, *PA* had the highest percentage of articles with a full replication archive, followed by *AJPS* at 85.6% and *IO* averaging 81.6%. At 27.9%, *APSR*—which expects but does not require authors to make materials available—had the fewest articles with full replication archives. Only 7% of articles provided either data or code but not both, which indicates that most authors who provide replication materials understand the importance of production and analytic transparency in addition to data availability.

As shown in table 2, authors more often fail to provide replication archives unless they are required to do so by journals. Those that require verification or mandatory provision generally have higher rates of replication availability than journals without such requirements, which lead authors to share more than their existing propensity to do so. It is interesting that there is no substantial difference in availability of full replication packages between journals that verify analyses and those that require replication materials be placed only in a journal's archives. That is, simpler policies that require less effort from journal editors appear to be as effective as more resource-intensive verification policies.

That is, simpler policies that require less effort from journal editors appear to be as effective as more resource-intensive verification policies.

Just as the availability of replication materials varies widely by journal, so also does the location of replication archives (table 3). Depositing data on journal-maintained databases is by far the most popular method of archiving, with 53.7% of replication materials housed in journal archives. This is encouraging yet unsurprising, given the mandatory policies of three of the journals examined. An additional 37 articles provided broken links to a replication archive on the journal's website, which demonstrates that archiving is not foolproof and highlights the need for persistent identifiers (Ishiyama 2014).

Beyond relying on journals to store replication materials, authors are turning to individual perpetual archives, such as personal dataverses or the Interuniversity Consortium for Political and Social Research. Only 7 of the 494 articles directed users to collect the data from websites such as the Bureau of Labor

Table 1

Replication Material Availability by Journal

Journal Name	Data and Code	Only Data	Only Code	Not Available	Total Articles
IO	81.6%	4.1%	4.1%	10.2%	49
AJPS	85.6%	3.4%	1.7%	9.3%	118
PA	94.4%	0%	3.7%	1.9%	54
BJPS	32.4%	16.2%	1.5%	50.0%	68
JOP	43.1%	3.6%	2.2%	51.1%	137
APSR	27.9%	4.4%	0%	67.6%	68

Table 2

Replication Material Availability by Policy

Replication Policy	Data and Code	Only Data	Only Code	Not Available	Total Articles
Verification	82.7%	3.4%	3.4%	10.3%	29
Mandatory Provision	88.4%	2.3%	2.3%	7.0%	172
Expected Provision	43.1%	7.6%	2.2%	47.1%	68
Not Required	27.9%	4.4%	0%	67.6%	68

Table 3

Replication Material Availability by Source

Source	Data and Code	Only Data	Only Code	Dead Links	Total Articles
Journal	75.5%	4.9%	2.5%	17.2%	212
Repository	68.9%	4.9%	1.6%	24.6%	61
Website Linked in Article	50.0%	9.5%	0%	40.5%	42
Searched for Website	81.0%	7.1%	3.6%	8.3%	84
Other External Website	18.2%	45.5%	0%	36.4%	11

Statistics, the Supreme Court Database, and the Correlates of War Project. Personal repositories, however, are more popular, with 16% of articles with replication materials stored in personal dataverses or other file-sharing sites. These archives are more durable than personal websites but are not without drawbacks. Similar to journal archives, 15 articles provided broken or password-protected links to data archives. If scholars provide their data through personal repositories, those archives must be maintained to remain accessible.

Personal websites remain popular for data storage, with 107 articles providing a link to an author's website. Of these, only 21 lead directly to a full replication archive, with an additional four linking to a dataset without a replication code. Most of the links to personal websites, however, are for the site's homepage rather than the replication archive.[4] Although this could be considered good practice if the website configuration changes frequently, 56% (40) of the homepages fail to provide pathways to replication materials on the site. In other words, readers are directed to search websites for data that are not there. Although this is a useful way to acquaint readers with other aspects of an author's work, it does not help them find what they most need. I have no reason to assume that this misdirection is intentional; there are many reasons why data may not be available on a personal website. Nevertheless, the replication materials are not being delivered in the way that they are promised.

Web searches for an author(s) were used when articles did not include any mention of replication archives or contained broken links, or when replication materials were not otherwise located. Those searches led to websites that contain replication materials for 77 articles, 68 of which contained both data and code. In other words, 26.4% of the total replication materials found were discovered through virtual digging; the need to search so thoroughly makes the replication processes more difficult than necessary.

More than 40% of links to websites included in articles were broken, which indicates that authors believe their replication materials are available when in reality they are not. This is particularly problematic for personal websites, especially when authors have changed institutions since publication of their articles. Moreover, these figures are based on recently published articles. As articles age, the likelihood of a "dead" or broken link increases. If scholars forgo dataverses and other durable archives, they must take extra care in maintaining their own websites.

As noted previously, some type of replication materials—data, code, or both—are available for 327 articles in the sample and full archives for 292. The strongest predictor of availability is whether a journal has a policy mandating that data and/or code be made publicly available at the time of publication (table 4). By requiring replication archives, it is 24 times more likely that any materials will be provided and 17 times more likely that a full replication package will be published. This echoes the claim of Gherghina and Katsanidou (2013, 337) that "[t]he most important element of a data availability policy is the extent it binds the authors." Likewise, journals with less stringent policies (i.e., *JOP*, *BJPS*, and *IO* before 2014) are more likely to have articles with replication archives than those that do not have replication requirements.

More than 40% of links to websites included in articles were broken, which indicates that authors believe their replication materials are available when in reality they are not.

The age of an article, measured as a count of the number of quarters since publication, does not have a significant effect on the likelihood of data sharing. It does not appear that authors find time to provide replication materials in the months following publication; neither has the discipline's recent focus on replication influenced the probability that newer articles will be published with replication materials. Rather, the degree of data access far more depends on a publishing outlet's policies.

Table 4

Logistic Regression of Factors Influencing Replication Material Availability

	Any Materials		Full Archive	
	Coefficient (Std. Error)	Odds Ratio	Coefficient (Std. Error)	Odds Ratio
Mandatory Provision	3.25* (0.30)	25.8	2.89* (0.23)	18.0
Expected Provision	0.86* (0.18)	2.3	0.68* (0.21)	2.0
Quarters Since Publication	−0.02 (0.08)	1.0	−0.04 (0.08)	1.0
Constant	−0.83 (0.34)	0.5	−0.81 (0.28)	0.4
N = 494				
Percent Correctly Predicted	71.05%		71.46%	

Notes: *p < 0.05. Standard errors clustered by journal.

DISCUSSION AND RECOMMENDATIONS

As with any collective action, diffusion of responsibility leads to shirking; the same is true for DA-RT. More than 33% of articles in the sample did not have publicly available replication materials; an additional 7% provided only some of the information needed for replication. To ensure greater cooperation, external enforcement mechanisms are necessary. The previous analyses confirm that the extent to which replication archives are provided is largely a function of journals requiring research transparency.

Although these replication policies are effective in increasing compliance, shifting the burden of research transparency to journals is costly. Whereas verification of the analyses presented in an article before publication is the "gold standard," it is unreasonable—and likely unnecessary—for all journals to implement such rigorous policies. Many journals lack editorial assistants, leaving the certification of results to editors. Considering the volume of submissions, verification of analyses is not feasible except at well-staffed journals.[5] In addition to concerns about efficient allocation of editorial time and effort, editors and staff may not have access to every program or add-on used in an analysis. Furthermore, simply because results can be verified using the data and code provided does not avoid situations in which the data contain serious errors or are somehow falsified.

Rather than verifying analyses before publication, journals should model their replication policies after journals such as *PA*, which requires that specific replication materials be uploaded to the journal's dataverse and cited in an article's references. This allows other interested scholars to verify and use the data and code and provides an opportunity for students to learn through replication (Janz 2015). It also relieves journals from the burden of duplicating results while still requiring that materials be made publicly available.

Even with mandatory provision policies, the compliance rate falls short of 100%. What are we to make of the approximately 20% of more substantive pieces that fail to fully comply with journal policies? The lack of availability may simply be an oversight on the part of authors or it may stem from a lack of appreciation for the importance of replication to the field as a whole. APSA, this journal, and others in the discipline stress the benefits of data access and replication, but the message has not reached everyone. Rather than devoting resources to the verification of results, journals can improve availability by certifying that authors have complied with replication policies before publication.

It is important to note that replication files for this analysis were downloaded but not opened or run and therefore may not be complete. By coding articles based on the availability rather than the integrity of the replication package, this article assesses only whether a minimum standard is being met. Journals should establish specific guidelines about the contents of a full replication archive (Altman and King 2007; Eubank 2014). The APSA section responsible for its journal also should maintain the journal's dataverse, alleviating work for overburdened editors. Last, there is a need for archives to be associated with articles through persistent identifiers rather than web links (Ishiyama 2014). In summary, data access cannot be the sole responsibility of individual researchers. Journals must take a more active role in building a culture of data sharing and ensuring research transparency.[6] ■

NOTES

1. See McCullough and McKitrick (2009) for a list of studies that failed to replicate results across a variety of fields.
2. Although arguably more difficult, there have been calls for qualitative studies to subscribe to the same standards at quantitative work (Elman and Kapiszewski 2014; Golden 1995).
3. *AJPS* requires citations to be in the first footnote, whereas *PA* includes data citations in the reference list.
4. An additional 17 articles contained broken links to personal websites.
5. *Political Science Research and Methods* is an example of a less-established journal that is able to require verification of analyses before publication.
6. Replication materials for this article are available at http://dx.doi.org/10.7910/DVN/5LJAMC.

REFERENCES

Altman, Micah, and Gary King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data." *D-Lib Magazine* 13 (3/4).

Anderson, Richard G., William H. Greene, B. D. McCullough, and H. D. Vinod. 2005. "The Role of Data and Program Code Archives in the Future of Economic Research." Working Paper 2005-014C, Federal Reserve Bank of St. Louis.

APSA. 2014. "Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors." Available at dartstatement.org.

Broockman, David, Joshua Kalla, and Peter Aronow. 2015. "Irregularities in LaCour (2014)." Available at http://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf.

Dafoe, Allan. 2014. "Science Deserves Better: The Imperative to Share Complete Replication Files." *PS: Political Science & Politics* 47 (1): 60–6.

Elman, Colin, and Diana Kapiszewski. 2014. "Data Access and Research Transparency in the Qualitative Tradition." *PS: Political Science & Politics* 47 (1): 43–7.

Eubank, Nicholas. 2014. "A Decade of Replications: Lessons from the *Quarterly Journal of Political Science*." *The Political Methodologist* 22 (1): 18–19.

Fowler, Linda L. 1995. "Replication as Regulation." *PS: Political Science & Politics* 28 (3): 478–81.

Gherghina, Sergiu, and Alexia Katsanidou. 2013. "Data Availability in Political Science Journals." *European Political Science* 12:333–49.

Gleditsch, Nils Petter, Claire Metelits, and Havard Strand. 2003. "Posting Your Data: Will You Be Scooped or Will You Be Famous." *International Studies Perspectives* 4 (1): 89–97.

Golden, Miriam A. 1995. "Replication and Non-Quantitative Research." *PS: Political Science & Politics* 28 (3): 481–3.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2013. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." Working Paper Series 332. Amherst: University of Massachusetts, Political Economic Research Institute.

Ishiyama, John. 2014. "Replication, Research Transparency, and Journal Publications: Individualism, Community Models, and the Future of Replication Studies." *PS: Political Science & Politics* 47 (1): 78–83.

Janz, Nicole. 2015. "Bringing the Gold Standard into the Classroom: Replication in University Teaching." *International Studies Perspectives* 1–16 (published online).

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28 (3): 443–59.

———. 2006. "Publication, Publication." *PS: Political Science & Politics* 39 (1): 119–25.

LaCour, Michael J., and Donald P. Green. 2014. "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality." *Science* 346 (2): 1366–9. (Retraction published May 28, 2015.)

Lupia, Arthur, and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency." *PS: Political Science & Politics* 47 (1): 19–42.

McCullough, Bruce D., and Ross McKitrick. 2009. *Check the Numbers: The Cases for Due Diligence in Policy Formation*. Vancouver, BC: Fraser Institute.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100 (2): 573–8.