Harvard Data Science Review • Issue 2.4, Fall 2020

Trust but Verify: How to Leverage Policies, Workflows, and Infrastructure to Ensure Computational Reproducibility in Publication

Craig Willis, Victoria Stodden

Published on: Dec 16, 2020

DOI: 10.1162/99608f92.25982dcf

ABSTRACT

This article distills findings from a qualitative study of seven reproducibility initiatives to enumerate nine key decision points for journals seeking to address concerns about the quality and rigor of computational research by expanding the peer review and publication process. We evaluate our guidance in light of the recent National Academies of Science, Engineering, and Medicine (NASEM, 2019) report on *Reproducibility and Replicability in Science* recommendation for journal reproducibility audits. We present 10 findings that clarify how reproducibility initiatives contend with a variety of social and technical factors, including significant gaps in editorial infrastructure and a lack of uniformity in how research artifacts are packaged for dissemination. We propose and define a novel concept of *assessable reproducible research artifacts* and point the way to an improved understanding of how changes to author incentives and dissemination requirements impact the quality, rigor, and trustworthiness of published computational research.

Keywords: reproducibility, reproducibility audits, reproducibility initiative, reproducibility policy, open data and code, peer review

---

1. Introduction

It is widely recognized that computation and data are of increasingly central importance for discoveries in a diverse set of fields. Across these fields, concerns are increasing about the rigor and trustworthiness of published results, arising from a lack of transparency and verifiability of computational methods (Anderson et al., 2008; Begley & Ellis, 2012; Chang & Li, 2017; Data Access and Research Transparency [DA-RT], 2015; Donoho et al., 2009; King, 1995; Krishnamurthi & Vitek, 2015; Peng et al., 2006; Yong, 2012). A recent National Academies of Science, Engineering, and Medicine (NASEM, 2019) consensus report called *Reproducibility and Replicability in Science* (one of us was a committee member) provides definitions of *reproducibility* and *replicability*, which we follow in this work. Specifically:

> *Reproducibility* is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "computational reproducibility," and the terms are used interchangeably in this report (p. 36).
>
> *Replicability* is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study (p. 36).

The report also defines *transparency* as "the extent to which researchers provide sufficient information to enable others to reproduce the results" (p. 51).[1]

The report goes on to recommend that "[j]ournals should consider ways to ensure computational reproducibility for publications that make claims based on computations, to the extent ethically and legally possible" while recognizing that this "presents technological and practical challenges for researchers and journals" (NASEM, 2019; See Appendix E for the full text of report recommendations). In addition to the definitions provided in the report, we define a *reproducibility initiative* as formal activities undertaken by journal editors, conference organizers, or related stakeholders to improve the transparency and reproducibility of computational research published via their venues through the adoption of new policies, workflows, and infrastructure. We define *computational research artifacts* as the packaged research artifacts (e.g. data sets, analysis code, workflows, and environment) generated and reviewed or verified as a result of these processes. Just as each of the reproducibility initiatives studied, we are concerned primarily with the concepts of computational reproducibility and transparency, rather than replicability.

While many have argued that sharing the data and code behind published research is a natural way to increase both rigor and trust (Anderson et al., 2008; Baggerly & Berry, 2011; Donoho et al., 2009; Donoho, 2010; King, 1995; Peng, 2011), today there are no widely accepted standards for how computational research artifacts—deemed necessary for computational reproducibility—should be shared or evaluated. We carried out a novel multiple-case analysis of seven reproducibility initiatives to address this question. The studied reproducibility initiatives, from the disciplines of political science, computer science, economics, statistics, and mathematics, provide concrete examples of how computational reproducibility and transparency can be assessed in practice. As suggested by the National Academies report recommendation mentioned previously, these initiatives face a variety of challenges, both social and technical. We comprehensively study the initiatives to better understand the many factors involved in the expansion of both the peer review and publication process to include new requirements for the assessment and dissemination of reproducible computational research artifacts. Based on our findings, we propose a general set of guidelines for new reproducibility initiatives along with an actionable and assessable definition for *reproducible research artifacts*.

This article is organized as follows. The methods section describes our experimental design and analysis approach. This is contextualized by a discussion of prior work in the next section. We then present 10 findings from our investigation, followed by a section discussing the nine key decision points we distill from our results. We then conclude with a discussion of open questions and future research directions and provide a short note on how reproducibility and replicability relate to this study in particular.

2. Experimental Design and Research Methods

We carry out a multiple-case analysis (Yin, 2017) of seven reproducibility initiatives designed to improve computational reproducibility across the fields of political science, computer science, economics, mathematics, and statistics. The initiatives were selected to represent both a broad range of disciplines as well as different requirements with respect to computational scale. To be representative of general approaches to reproducibility review, each initiative was required to have established policies and workflows for a minimum of three years.[2] The initiatives are: the *American Journal of Political Science* (*AJPS*) (Christian et al., 2018; Jacoby et al., 2017); the ACM/IEEE *Supercomputing* (SC) conference;[3] the American Economic Association (AEA) (Vilhuber, 2019); the *Biostatistics* journal (Peng, 2009); the *Information Systems* journal (*IS*) (Chirigati, Capone et al., 2016); the *Journal of the American Statistical Association-Applications and Case Studies* (*JASA-ACS*) (Fuentes, 2016); and the ACM *Transactions on Mathematical Software* (*TOMS*) (Heroux, 2015). The *AJPS*, AEA, *JASA-ACS*, and *Biostatistics* journals tend to publish research considered in the 'long tail' or small-scale data analysis that tends to leverage statistical methods and tools, whereas *SC*, *IS*, and *TOMS* produce research artifacts associated with high-performance computing environments. The AEA and *AJPS* communities have histories of reproducibility discussions dating back to the 1980s and 1990s (Dewald et al., 1986; King, 1995) and similarly the computationally focused communities have histories of discussion for almost as long (Claerbout & Karrenbach, 1992), however, these various discussions generally did not cut across disciplines.

Case profiles were developed from three primary sources of evidence including:

- Interviews with key informants (n = 17 editors, reviewers, verifiers, and curators)[4] (see Appendix A for the interview protocol).
- Publicly available documents from each initiative, including policies, guidelines, workflows, editorials, and editor reports (see Appendix B for a complete list).
- A representative sample of (n = 27) artifacts that have been reviewed or verified through these initiatives, including packages of code and data, reproducibility papers, and reports (see Appendix C for a complete list).

All data were collected between October 2019 and March 2020 and reflect the state of each initiative at that time. Qualitative analysis was conducted in two phases. The first phase focused on individual case analysis for case profile development. The second phase focused on cross-case analysis of the seven initiatives. Qualitative coding and analysis were applied following the method described by Schreier (2012). Qualitative code development was informed by a preliminary literature review in the areas of experimental reproducibility (Radder, 1996), computational reproducibility (Freire et al., 2016), and knowledge infrastructures (Edwards, 2010; Star & Griesemer, 1989; Star & Ruhleder, 1996) and refined throughout the process. Open coding was used to identify themes and concepts and a final set of codes selected for focused coding. High-level codebooks are included in Appendix D. Case profiles were compared to identify key or common factors that contribute to operational decisions across the initiatives as well as to identify possible explanations for the observed similarities or differences. A summary of the case profile structure is also included in Appendix D. Complete case descriptions, including details of each initiative's operational workflow along with other study details, are provided in (Willis, 2020a, 2020b). All quotations reported below from the interviews are identified by "(Interviewee #-#)" which map to study participant identifiers. Interview transcripts are confidential.

3. Prior Work

To our knowledge, no one has studied publication reproducibility initiatives to understand the factors involved in their implementation. Prior work in this area has generally focused on tools and methods for disseminating computational research; studies of the extent of irreproducibility within disciplines; and the incentives and costs associated with the production of computationally reproducible research. This earlier work informs our investigation as we focus on how communities operationalize the assessment of reproducibility through the peer review process.

What we call today 'computational reproducibility' has its origins in four different traditions that present distinct views of what it means to share the data and code behind published research. First, the early efforts in computer science, mathematics, and statistics toward the review and distribution of high-quality scientific software libraries (Hopkins, 2009; LeVeque, 2006). Second, the 'replication standard' movement of the 1980s and 1990s in political science and economics, exemplified by the work of King (King, 1995). King proposed that authors should share the code and data behind published political science research for the evaluation and ultimately replication of their work. Third, the 'reproducible research' movement started in the early 1990s by geoscientists Claerbout and Karrenbach (1992) and more generally adopted in statistics (Buckheit & Donoho, 1995; Peng, 2009) and signal processing (Kovacevic, 2007; LeVeque, 2006). They first introduced the phrase "reproducible research" (Barba, 2018; Claerbout & Karrenbach, 1992) to describe their vision of "merging publication with its underlying computational analysis." They envisioned a system where the local software environment, data, and analysis code could be used to reproduce the publication, including tables and figures, by "pressing a single button" and went so far as to claim that the "[j]udgement of the reproducibility of computationally oriented research no longer requires an expert—a clerk can do it" (Plesser, 2018). Finally, the 'repeatability' movement in computer science, started in the databases community (Manolescu et al., 2008). Each of these antecedents presents an alternative view into what it means to share the data and code behind published research that underlie the reproducibility initiatives of today.

The computational reproducibility movement of today has been fueled by the growing perception of a "crisis" in research reproducibility and credibility across the sciences (Baker, 2016; Begley & Ellis, 2012; Fanelli, 2018; Open Science Collaboration, 2012; Spiegelhalter, 2017). With the emergence of the 'reproducibility crisis' narrative in 2005, many communities began looking for ways improve the rigor of published research. Proposed solutions have included improvements to study design and power (Ioannidis, 2005), study preregistration (Open Science Collaboration, 2012), changes in practice related to statistical significance (Wasserstein & Lazar, 2016), and increased research transparency (DA-RT, 2015). For fields and subfields with a focus on computational methods, the idea of publishing reproducible computational research has increasingly been seen as a way to promote transparency, to increase confidence in published work, and to quickly identify and correct sources of error.

The urgency of the problem of computational reproducibility has more recently been highlighted by multiple attempts across disciplines to reproduce results reported in the literature. There have been many such studies and we point to a few for context. For example, a study of reproducibility in computer science research found that 32.1% of the 20 experiments could be reproduced when not communicating with the authors and 48.3% when communicating with the authors (Collberg & Proebsting, 2016). In a recent study in economics, fewer than half of the 67 articles studied could be reproduced with the assistance of authors (Chang & Li, 2017). In computational physics, no articles were fully reproduced out of 306 studied (Stodden, Krafczyk, & Bhaskar, 2018). Finally, a study of articles published in *Science* found that only 26% were computationally reproducible (Stodden, Seiler, & Ma, 2018). The results of these reproducibility studies have led communities to consider the adoption of methods to ensure the reproducibility of published research, including those studied in our work.

Reproducibility initiatives can be found across the sciences in fields as diverse as political science (Alvarez et al., 2018; Eubank, 2016; Jacoby et al., 2017), economics (Vilhuber, 2018), computer science (Fursin & Dubach, 2014; Krishnamurthi, 2013; Manolescu et al., 2008), mathematics (Heroux, 2015), and statistics (Fuentes, 2016). Several of these initiatives represent the latest evolution of policies over a period of years or decades. Vilhuber (2018) summarizes the history of reproducibility in economics where, over a period of decades, repeated attempts to reproduce the results of computational research (e.g., Chang & Li, 2017; Dewald et al., 1986; McCullough & Vinod, 2003) have led to even stricter publication policies (e.g., Ashenfelter et al., 1986; Bernanke, 2004; Vilhuber, 2019). Similar examples can be found in political science (e.g., Jacoby et al., 2017; Meier, 1995; Wilson, 2012). Our work seeks to apply the lessons learned from these initiatives.

Efforts to improve computational reproducibility have resulted in a remarkable amount of technical infrastructure designed to support the creation, publication, and distribution of computationally reproducible research artifacts. Konkol et al. (2020) present a comparison of the technical features of many tools from the perspective of authors. Our work complements theirs by focusing instead on the social and technical infrastructure used in the assessment of reproducibility by publication initiatives. For computational reproducibility, the technical dimensions associated with tools and infrastructure are important; however, expanding the peer review process also has important social and organizational dimensions.

More broadly, our work is related to the literature on incentives and costs associated with the production of computationally reproducible research. The early work of Dewald et al. (1986) on policy changes at the *Journal of Money, Credit, and Banking* (*JMCB*) found that authors were most likely to provide complete computational artifacts after an article had been accepted but prior to publication. Building on this work, Mirowsky & Sklivas (1991) conclude that improving reproducibility in the field of economics would require editors to increase the information requirements on authors or find alternatives to current incentives. Feigenbaum & Levy (1993) find that there are powerful disincentives to authors to provide reproducible research artifacts as long as irreproducibility is not factored into publication or promotion. In a more recent discussion in the field of psychology, Nosek et al. (2012) argue for reducing barriers to publication to shift away from publication incentives and advocate for the use of checklists in place of stricter verifications. Our work supports the conclusions of Mirowsky and Sklivas (1991) and Feigenbaum and Levy (1993) while also leaving open those of Nosek et al. (2012) for future work.

4. Findings: The Importance of Editorial Roles and the Interpretation of Reproducibility

The reproducibility initiatives have resulted in new publication policies and workflows that expand the peer review and publication process beyond its framework of article publications and introduce new requirements for the sharing and assessment of the code, data, and computational workflows behind claims made in published manuscripts. While outwardly the initiatives have similar goals, they differ widely with respect to policy mandates, what is reviewed, who conducts the review, and how reviewers are incentivized. We also find that several initiatives are constrained by existing editorial infrastructure as well as access to the computational infrastructure required for review. These differences reflect some of the “technological and practical challenges” mentioned in NASEM Recommendation 6-4 (see also Appendix E). In the following sections we distill the new roles of editors and editorial policies, the importance of defining reproducibility, and the influence of supporting infrastructure on initiative success.

4.1. Editorial Roles, Mandates, and Policies

Our first major finding is that *initiatives introduce specific new editorial roles and policies to enable their efforts*, summarized in Table 1. These new roles are responsible for shepherding the reproducibility process and are named variously Data Editor, Associate Editor for Reproducibility, Reproducibility Editor, Reproducibility Chair, and Replicated Computational Results Editor. Additional titles, such as Curator and Verifier, refer to new positions that carry out specific reproduction, transparency, and preservation activities. In almost all cases, new roles and positions were created instead of assigning new duties to an established editorial position.

Table 1. Initiative, Organization, Roles, and Policies (as of February 2020).

Initiative	Organization	Roles	Policy
AEA	Centralized (LDI/Cornell)	Data Editor and verifiers	https://www.aeaweb.org/journals/policies/data-code
AJPS	Centralized (Odum/UNC)	Curators and verifiers	https://ajps.org/ajps-verification-policy
Biostatistics	Decentralized	Associate Editor for Reproducibility	https://academic.oup.com/biostatistics/pages/General_Instructions
IS	Decentralized	Reproducibility Editor	https://www.elsevier.com/journals/information-systems/0306-4379/guide-for-authors
JASA-ACS	Decentralized	Associate Editor for Reproducibility	https://jasa-acs.github.io/repro-guide/pages/author-guidelines
SC	Decentralized	Reproducibility Chair	https://sc19.supercomputing.org/submit/reproducibility-initiative
TOMS	Decentralized	Replicated Computational Results (RCR) Editor	https://dl.acm.org/journal/toms/replicated-computational-results

Note. AEA = American Economic Association; *JASA-ACS =Journal of the American Statistical Association-Applications and Case Studies*; *Biostatistics* journal; *IS = Information Systems* journal; *SC = Supercomputing* conference; *TOMS = Transactions on Mathematical Software*.

Six of the seven initiatives created new editorial roles with specific responsibility for the assessment process, either directly or through the recruitment or supervision of reviewers. In the outstanding case (*AJPS*), responsibility for the assessment process was given to staff at the H. W. Odum Institute for Research in Social Science at University of North Carolina at Chapel Hill (UNC). The Odum staff rely on a journal managing editor to serve as liaison.

As in conventional peer review, in most cases the reproducibility assessment process is decentralized and managed by editors and reviewers at academic institutions. However, in two initiatives (AEA, *AJPS*), assessment is centralized in a research center at a single host institution. As will be discussed later, centralization of the assessment process allows initiatives to leverage infrastructure and additional expertise provided by their host institution and can especially assist with computational execution of author-submitted artifacts.

Our second finding is that *an essential component of each reproducibility initiative is a clearly articulated policy that is made available to authors*. In the initiatives studied here, these policies are generally accompanied by guidelines, checklists, and workflows that define their operationalization for both authors and reviewers. Each initiative's policy document is found at the links in the last columns of Table 1. As will be discussed, differences in initiative organization are reflected in how these policies are realized.

Our third finding is that *the strength of reproducibility mandates arises from community readiness and initiative scale*. Initiative policies determine whether the assessment process is mandatory or voluntary. As can be seen in Table 2, there are three types of policy mandates across the seven initiatives: all manuscripts are subject to assessment (mandatory), authors agree to the assessment process (opt-in), or the editors invite authors to participate (invited).

Table 2. Initiative Mandate, Role, and Number of Assessed Artifacts (as of February 2020).

Initiative	Year	Mandate	Artifacts	When assessed
AEA	2019– [a]	Mandatory	> 200	Conditional accept [b]

AJPS	2015– [c]	Mandatory	> 200	Conditional accept
Biostatistics	2009–2011[d]	Opt-in	< 5	Conditional accept
IS	2016–	Invited	< 5	Post-publication
JASA-ACS	2016–	Mandatory	> 50	Conditional accept
SC	2015–	Mandatory	> 50	Conditional accept
TOMS	2015–	Opt-in	< 5	Conditional accept

[a] The *American Economic Review's* (*AER)* first Data Availability policy dates from 1986, with the current AEA-wide policy from 2019.

[b] AEA now requires materials to be provided prior to paper acceptance, but the assessment still occurs after acceptance (Vilhuber et al., 2020).

[c] *AJPS* implemented its first Replication policy in 1994, but the verification initiative began in 2015.

[d] *Biostatistics* has not had a reported reproduction since 2011.

Note. See Table 1 for abbreviations.

Mandatory assessment indicates that publication leadership is confident that the community will accept the additional burden without having a significant impact on submission rates or measures of impact.[5] On the other hand, through opt-in policies authors voluntarily submit to the additional review and therefore initiatives can reduce the risk of pushback or a negative impact on submissions. Opt-in policies also allow initiatives to scale up as demand increases within the community. Invitation-only policies similarly allow initiatives to control the number of reviews while also being selective about papers reviewed. As can be seen by the number of artifacts reviewed by each initiative in Table 2, opt-in policies have remarkably low participation rates even after several years of initiative activity. One editor noted, "I guess maybe that was predictable. Not many people would voluntarily submit to this just for the hassle alone [...]" (Interviewee 4-1). Initiatives with mandatory assessment processes have a history of addressing community buy-in and making operational preparations for the scale of the review process.

4.2. Interpreting Reproducibility: When, of What, and by Whom?

Our next finding is that *reproducibility review always occurs post-acceptance*, as shown in the rightmost column of Table 2. In six of the initiatives, papers are conditionally accepted pending successful reproducibility review. The assessment is a condition of publication only and has no bearing on the manuscript acceptance decision. The *IS* initiative is unique in that the invited reproducibility paper is based on an already published work. The use of conditional- and post-acceptance review raises questions about what happens when results cannot be reproduced, as most of the initiatives do not have specific policy provisions for handling instances of nonreproducibility aside from publication delay. *TOMS* is the only initiative with a stated policy on nonreproducibility:

> RCR [Replicated Computational Results] Review Failure: There is some risk now and in the future that RCR efforts will fail. In this case, we must acknowledge that the manuscript is not ready for publication with the presented results. During the introductory phase, the EiC will personally manage this situation if it occurs and will work with the authors to avoid rejecting the manuscript outright. As the RCR initiative matures, we anticipate that failed RCR reviews would constitute grounds for returning the manuscript back to the authors for revision, or for rejection if concerns were serious.

This provision suggests a rationale behind post-acceptance review. First, conditional acceptance and the absence of policy provisions for irreproducibility indicate that initiatives generally expect that authors of accepted papers will either be able to provide reproducible artifacts or revise the manuscript without fundamental changes to their findings. Second, the reproducibility assessment process is presented as a supportive activity. One initiative chair noted:

> In fact, calling [the role] "reviewer" is not technically the best word for it. It was more an advisor. They would work with the authors to try to improve the quality of their artifact and get it to a point where we felt that all the hardware, software and data had been fully described in a way that a third party would understand the experimental setup. (Interviewee 4-7)

As suggested in the RCR provision, as these initiatives mature, the reproducibility assessment process may have more bearing on acceptance decisions in the future. This is not the case today in the studied initiatives, where reproducibility review occurs post-acceptance and is only a condition of publication.

We also find that *reproducibility initiatives must set policy to decide* what *to reproduce and* by whom. The seven initiatives differ widely in how the reproducibility assessment process is operationalized as well as how reviewers are incentivized. Table 3 summarizes these characteristics across the initiatives.

Table 3. Summary of Initiative Assessment Characteristics (as of February 2020).

Initiative	What is assessed	Who assesses	Incentive
AEA	Data and attempted full reproduction	Supervised graduate or undergraduate student	Paid
AJPS	Data and full reproduction	Curator and supervised advanced graduate student or professional statistician	Paid
Biostatistics	Full reproduction	Associate editor	Position
IS	Full reproduction with extension	Peer	Publication
JASA-ACS	Materials review (reproduction optional)	Associate editor	Position
SC	Materials only	Peer	Voluntary
TOMS	Full reproduction	Expert practitioner	Publication
Note. See Table 1 for abbreviations.			

When considering *what* is assessed during the review process, there is an important distinction between 'reprodu*cibility*'—assessing the possibility of reproduction—and 'reprodu*ction*' or actually reproducing the results (Radder, 1996). Each of the initiatives directly operationalizes these concepts through their defined workflows and represent different approaches to reproducibility assessment. *Materials only*, shown in Column 2 of Table 3, requires that reviewers only assess author provided materials without any attempt at reproduction (i.e., running the code). *Partial reproduction* occurs when reviewers reproduce only a subset of results. This is reflected, for example, in the AEA policy statement that code will be re-executed "when feasible." *Full reproduction* occurs when reviewers are required to re-execute all code and assess results as compared to the published manuscript. *Full reproduction with extension* includes full reproduction and requires that reviewers attempt to extend the submitted work, for example through changes to parameters, input data, or input conditions.

The ability of an initiative to mandate full or partial reproductions is also related to both initiative organization and the scale and complexity of the computational elements of submitted manuscripts. The two initiatives that mandate full or partial reproductions (AEA and *AJPS*) are centrally organized and rely on the computational infrastructure provided by their host institutions (see Table 1 and Table

4). These initiatives also tend to have fewer papers that rely on highly computationally intensive methods, so reproduction tends to be tractable. The decentralized initiatives rely on reviewer (or, less commonly, author) computational infrastructure (see Table 4) and therefore access to the resources required to conduct a full reproduction may pose a problem. For those initiatives where scale and complexity are generally high, full reproductions may not be possible and alternative modes of assessment are required. When discussing why materials-only review was selected over reproductions, one editor noted:

> [T]he challenge that we felt was that, a large fraction of the papers that we get to [initiative] use fairly computationally intensive methods. This is going to run for eight hours or requires a cluster or whatever and we just didn't feel that it was going to be feasible to do that for every paper and in any reasonable amount of time. (Interviewee 3-2)

With respect to who conducts the assessment, the seven initiatives represent three broad approaches: peers, expert practitioners, or students under the guidance of another responsible party. The two initiatives that rely on students are centrally organized, conduct reproductions, and have well-documented workflows. The initiatives that rely on peers or expert practitioners are decentralized, less likely to conduct full reproductions, and tend to trust the reviewer's expertise in the conduct of their assessment.

Finally, initiatives have also had to implement new incentives for reviewers. There are four models of incentives: reviewers are compensated financially (paid); the editorial position itself is the incentive (position); the reviewer gains a publication[6] (publication); or the reviewer volunteers. Mandatory reproductions, such as those in the *AJPS* and AEA initiatives, rely on a combination of financial incentives and experience gained by students.

Table 4. Summary of Initiative Infrastructure (as of February 2020).

Initiative	Publisher	Editorial Software	Adaptation	Compute Resources
AEA	AEA	ScholarOne	Custom database	Cornell
AJPS	Wiley	Editorial Manager	Custom database	UNC
Biostatistics	Oxford	ScholarOne	None	Reviewer
IS	Elsevier	Editorial Manager	Companion publication	Reviewer
JASA-ACS	Taylor & Francis	ScholarOne	None	Reviewer

SC	ACM	Linklings	Custom database	Reviewer
TOMS	ACM	ScholarOne	Companion publication	Reviewer or author
Note. See Table 1 for abbreviations.				

4.3. Common Initiative Requirements and Infrastructure

Our sixth finding shows there are *common requirements for reproducibility across initiatives*, for example, access to software artifacts used in generation of results and exposure of details of the computational environment. Most initiatives also require: documentation of computational workflow, access to data used in generation of results, long-term accessibility of artifacts, provenance of results, data licensing, provisions for proprietary and confidential data, and details of the experimental context (see Table 5).

Table 5. Core factors in reproducibility assessment.

Requirement	Initiatives
Access to software artifacts used in generation of results	All
Details of the computational environment	All
Documentation of computational workflow	All except *TOMS*
Access to data used in generation of results	All except *TOMS*
Long-term accessibility of artifacts	*AEA, AJPS, JASA-ACS, IS*
Provenance of results	*AEA, AJPS, JASA-ACS*
Data licensing	AEA, *Biostatistics, IS, JASA*
Provisions for proprietary and confidential data	*AEA, AJPS, IS, SC*
Details of the experimental context	*AEA, IS, SC*
Note. See Table 1 for abbreviations.	

Initiative policies define the types of information required of authors to comply with the assessment process. Some of this information is directly related to the assessment of computational reproducibility while other information is required for understandability and reusability of provided artifacts.

Our seventh finding is that *reproducibility initiatives rely on established repositories for artifact preservation, stewardship, and long-term access*. As shown in Table 6, the initiatives recommend or require a variety of repositories that may be discipline dependent. All but one initiative require researchers to deposit artifacts in an archival repository. Two initiatives (*AJPS*, *IS*) require authors to deposit materials in initiative-specific repositories (Dataverse, Mendeley Data). One initiative (AEA) encourages deposit in a specific repository (OpenICSPR) but accepts submissions from other approved archives. Two initiatives (*Biostatistics*, *SC*) encourage the use of general-purpose repositories (e.g., Zenodo and Figshare). One initiative (*JASA-ACS*) requires submission of supplemental information via the publisher, which is made available via Figshare and Github. The final initiative (*TOMS*) only requires that authors make materials available for the review process and offers multiple different approaches, including guest access to remote systems.

Table 6. Platforms Required or Recommended by Each Initiative.

Initiative	Recommended and required dissemination platforms
AEA	OpenICSPR
AJPS	Dataverse
Biostatistics	Zenodo, Figshare
IS	Mendeley Data
JASA-ACS	Dataverse, Dryad, Zenodo
SC	Any DOI-minting repository
TOMS	Not specified
Note. See Table 1 for abbreviations.	

Despite more than two decades of development of tools and infrastructure in support of computational reproducibility, current initiatives rely on few. Only two initiatives provide any guidance on the use of specific packaging formats or reproducibility tools. *Biostatistics* encourages the use of literate programming environments. *IS* recommends packaging the environment via virtual machine images, Docker images, or using ReproZip.[7] Otherwise, research repositories are the central reproducibility infrastructure required by the initiatives.

Our eighth finding is that *current editorial technical infrastructure is insufficient for computational reproducibility by journals and must be adapted or alternate mechanisms put in place*. Editorial management platforms are central to the peer review and publication process. Software such as Scholar One or Editorial Manager are widely used to track and manage the communication between editors, reviewers, and authors throughout the peer review process. However, these systems are not designed to support the reproducibility review process, which focuses primarily on computational research artifacts, is generally not part of conventional peer review, and may require access to computational resources and licenses. As a result, initiatives have had to address or work around these limitations. Table 4 summarizes the key infrastructure required for reproducibility review across the seven initiatives.

Three initiatives (AEA, *AJPS*, and SC) have developed custom tools to manage the reproducibility review process. This includes handling reviewer assignment, tracking the review process, capturing versions of artifacts over time, and managing the reproducibility reports. These custom databases are often used in conjunction with the repository systems listed in Table 6. Two initiatives (*IS*, *TOMS*) treat the reproducibility review as a companion publication, leveraging existing paper-centric editorial infrastructure.

For those initiatives that conduct actual reproductions, access to computational resources and licenses are essential. As can be seen in the fourth column of Table 4, these are typically provided by the initiative host institution or depend on resources available to reviewers at their local institutions if initiatives are decentralized. Access to computational resources is a factor in whether initiatives can mandate full reproductions for all manuscripts.

Table 7. Summary of Initiative Metrics (as of February 2020).

Metric	Description

Impact factor	Measures of journal impact such as the Journal Citation Reports (JCR) 2 or 5 year, Google h-Index, SNIP. Used largely anecdotally to gauge whether policy changes are correlated with positive or negative changes in impact over time.
Manuscript submissions	Number of manuscripts submitted. Used to gauge whether policy changes are affecting submission rates.
Artifact resubmissions	Number of times research artifacts are resubmitted during the assessment process. Used as an indicator of author errors and effort required by assessors.
Assessment duration	How long the assessment process takes for a manuscript.
Author response time	How long it takes for the author to make corrections and resubmit materials.
Publication delay	Number of days added by the review to the publication process.
Assessment cost	Per-manuscript cost of assessment, generally determined by assessment time.

Our ninth finding is that the *reproducibility initiatives use operational metrics to measure policy effect*. Table 7 summarizes some of the metrics used. These include impact measures, such as journal impact factor, and operational measures including the number of artifact resubmissions,[8] review time taken, and time added to the publication process. Manuscript submission rates are typically used to determine whether policy changes are having a negative impact on submissions. Two cases report monitoring the Journal Citation Reports (JCR) impact factor (*AJPS*, *Biostatistics*). Three cases report monitoring manuscript submission and acceptance rates (*AJPS*, *JASA-ACS*, *SC*). *AJPS* and AEA additionally report publication delays caused by the review process and the number of artifact resubmissions. The *AJPS* reports the average time required for review (Christian et al., 2018), which can be converted to a cost estimate.[9] Notably, while initiatives typically do capture information about the errors encountered during the review process, this information is not currently part of recorded metrics.

Our final finding is that *there is no common standard for the description and packaging of reproducible computational research artifacts*. Initiatives have defined their own requirements and approaches for authors, generally relying on research data repository infrastructure as described in Table 6.

We summarize our findings in Table 8.

Table 8. Findings From a Multiple Case Study of Seven Reproducibility Initiatives.

Results from the study of seven reproducibility initiatives
1. Reproducibility initiatives introduce specific new editorial roles and policies.
2. Each initiative makes a clearly articulated policy available to authors.
3. Mandate strength arises from community readiness and initiative scale.
4. Reproducibility review occurs post-acceptance.
5. Reproducibility initiatives set policy to decide *what* to reproduce and *by whom*.
6. There are common requirements for reproducibility across initiatives.
7. Reproducibility initiatives rely on established repositories for artifact preservation and access.
8. Editorial infrastructure must be adapted or alternate mechanisms put in place.
9. Reproducibility initiatives use operational metrics to measure policy effect.
10. Lack of standards for the description and packaging of reproducible research artifacts.
Note. For each of the findings both social and technical factors are at play in initiative response.

5. Guidelines and Recommendations

In this section, we interrogate our findings to distill 10 decision points to guide new reproducibility initiatives. The audience for these decision points are those stakeholders involved in the creation of new reproducibility initiatives, typically journal, conference, or association leadership. The investigated initiatives were largely undertaken by journal lead editors and conference organizers with association and community support, sometimes in collaboration with leadership in research data repository infrastructure. We cannot stress enough the importance of the role of journal editors and conference organizers in the success of these initiatives. We close the section by considering the future of reproducibility initiatives as they increasingly leverage software tools and infrastructure.

5.1. A Guide to Implementing a Reproducibility Initiative

Any reproducibility initiative will face the same operational decisions under many of the same constraints as the ones studied in this work. As we have seen, the initiatives share some broad similarities but differ widely in their implementations: organizational structures, mandates; the scope and depth of review as well as who performs the review; how they are incentivized; and what resources are required or available to complete the task. The 10 decision points presented are key to implementing the NASEM recommendation 6-4 for prepublication reproducibility review, and for a new reproducibility initiative to consider.

1. Assess where you are in terms of community readiness.

Requiring computational reproducibility imposes substantial changes for the research community as well as associated infrastructure, including the introduction of new organizational structures, editorial roles, policies, as well as researcher, journal, and reviewer workflows. Publisher support is important for establishing workflow changes and communication with authors. Initiatives rely on repository services, so repository readiness, whether domain-specific or general purpose, is also important.

Regardless of specific operational decisions, making this type of change will require a significant investment of time, leadership, and a commitment to carry through the vision. Assessments of community readiness can be conducted informally among publication leadership, through community surveys (e.g., Ferro & Kelly, 2018), or symposia dedicated to the discussion of policy changes (e.g., *PS: Political Science & Politics* 28:3 and *Biostatistics* 11:3). These open discussions can provide a diverse set of viewpoints on proposed policy changes.

As will be discussed following, initiative leaders must determine and clearly communicate to the community who will conduct the review, to what depth, and how reviewers are incentivized. They must work within the social norms and organizations already in place as well as within the constraints of existing editorial and publishing infrastructure while possibly introducing new infrastructure for the review and dissemination of computational artifacts. These operational decisions will codify what they mean by reproducibility.

2. Determine the strength of your mandate.

Mandatory policies ensure that all papers are treated equally but require an organization and infrastructure capable of efficiently handling the reproducibility review process. Opt-in policies, on the other hand, may be effective for piloting and ease scaling up as demand increases. However, it should be recognized that opt-in policies risk selectivity-bias since those who participate are already confident in the reproducibility of their work (Feigenbaum & Levy, 1993). A stronger journal mandate provides a critical incentive for authors who are otherwise disincentivized to provide reproducible computational artifacts (Feigenbaum & Levy, 1993; Mirowski & Sklivas, 1991). Even in the absence of full reproductions, mandatory policies should be prioritized to ensure that all papers receive the same scrutiny. Opt-in policies should be considered for temporary scaling or piloting purposes.

3. Determine what will be reproduced and by whom.

Full or partial reproductions by or on behalf of the journal ensure reproducibility and reduce errors. As first observed by Dewald et al. (1986) and reconfirmed by both the earlier SIGMOD (Special Interest Group on Management of Data) repeatability experiment (Bonnet et al., 2011; Manegold et al., 2010; Manolescu et al., 2008) and current *AJPS* initiative (Jacoby et al., 2017), materials provided by authors are generally incomplete and contain inadvertent errors. Assessment of reproducibility without reproduction will likely result in artifacts that contain oversights and errors. While full or partial reproductions present the best approach to ensuring reproducibility, this may not be possible due to the scale or complexity of reported research. In these cases, assessment of reproducibility or transparency through the inspection of materials may be the only feasible option. For initiatives that decide to implement full or partial reproductions, this still may not be tractable in some cases. In the event of large-scale or long-running computations, initiatives should consider the use of alternate methods such as reduction tests (Krafczyk et al., 2019) or metacomputations (Heroux, 2019) to demonstrate that the published code and data are working properly, or the use of computational provenance information (McPhillips et al., 2019) to show that the provided code and data were actually used in the generation of results.

4. Select a review structure: Centralized, decentralized, or hybrid?

Manuscript peer review is generally a decentralized process, engaging editors and reviewers at a variety of academic institutions. The advantages this model provides are scalable access to required expertise and familiarity to any research community. Centralized operations, such as the AEA and *AJPS* however, can rely on resources available to a discipline-specific research center, which may include a pool of students or practitioners with access to institutional computational resources. These initiatives also dedicate funding to support the reproducibility review process. We can also envision a hybrid model where decentralized reviewers leverage centralized human and/or computational infrastructure. Reviewers might have access to a pool of students or computational resources to conduct the reproducibility review without needing to be part of the same central organization. The select review structure will likely shape further decisions discussed following, including who conducts the review and how they are incentivized.

5. Select review management infrastructure.

Depending on the scope of review, reproducibility assessment may require integration with existing editorial and publishing platforms; the identification and selection of suitable research repositories; and, in the case of reproductions, access to computational resources. Any new initiative should consider whether to require or recommend use of a specific repository. Today, mature archival repository systems such as Dataverse, OpenICSPR, and Zenodo are widely available and have been demonstrated to be useful for the assessment process. However, both editorial and repository systems still lack capabilities necessary to track the artifact review process or conduct reproductions. Reproducibility initiatives develop their own tracking tools, in some cases through services such as Google Docs or other infrastructure. AEA defined a Jira[10] workflow and *AJPS* and *SC* have chosen to define customized relational databases for reproducibility review management. Infrastructure for conducting reproductions is discussed further below.

6. Decide whether to engage students in the review process.

In each of the studied initiatives, reproducibility review does not require expert knowledge of the research domain. Today, reviewers are not assessing the correctness of computations or considering the theoretical implications of the research. In the case of full reproductions, reproducibility review requires technical skills to configure, re-execute, and troubleshoot computational workflows. If deep domain knowledge is not required, then students may stand to gain significantly from the experience and exposure to new computational research methods, providing a useful incentive to conduct reproductions. However, the use of students presents additional challenges in terms of accountability and mentorship. Initiatives leveraging students must manage the work that needs to be done and confirm its quality. The *AJPS* requires its student verifiers to sign a nondisclosure agreement. However, the stakes may otherwise be low since the task before them is only confirming or disconfirming their ability to re-execute computations.

7. Define your policy, guidelines, and workflow.

We consider the top five broad requirements in Table 5 to be common across all initiatives and essential for the assessment of computational reproducibility. These include documentation of the computational workflow and access to precise versions of all software and data used in the generation of results, sufficient details of the computational environment to support third-party reproductions, and long-term accessibility of artifacts through archival repository infrastructure. Results provenance —documentation of the relationship between code/data and results—is important in the assessment of reproducibility primarily in the absence of an actual reproduction, given all other artifacts are provided. Artifact licensing is crucial for dissemination and reuse, although author permission can be easily obtained for the reproducibility assessment process. Similarly, for computational reproducibility, additional information about the experimental context beyond the provided workflow is not essential.

While the studied initiatives have developed their own policies, guidelines, and workflows, because of the many similarities across the initiatives, we believe that any new initiative should be able to adopt or adapt from these for their own use. We envision a collaboratively maintained set of resources based on the work of current initiatives that can easily be repurposed for new initiatives.[11] As part of ongoing work for a new initiative in the area of computational biology, we have developed and made available a preliminary checklist based on the policies of the studied initiatives.[12]

8. Choose appropriate operational metrics.

The metrics listed in Table 7 represent some of the information available to initiative leaders to track the effect of policy changes on publication operations as well as the efficiency of the assessment process itself. Impact factors and submission rates are often already tracked as part of journal operations and can be used to determine new policy effects. Examples can be found in the editor reports of the *AJPS* and AEA. The number of resubmissions, review time taken, and time added to the publication process are common measures of the assessment process itself. Others have suggested tracking the number and types of errors encountered (Alvarez et al., 2018; Hamermesh, 2007). Monitoring this information over time can also be used to measure operational changes in the assessment process itself. Additional metrics may be available from selected infrastructure. For example, publisher platforms may report the number of views, downloads, or citations of papers. Similarly, repository platforms may report the number of views, downloads, or citations of artifact packages or data sets.

9. Select reproduction infrastructure (if applicable).

A challenge for several of the initiatives is how to assess reproducibility when the research relies on private/protected resources or requires large-scale computational resources. The initiatives recognize that private or protected data, software, and hardware can affect both the reproducibility review and any subsequent reproduction or reuse scenarios. In one solution, the AEA requires authors to provide detailed "access protocols"—detailed descriptions of how a reviewer with appropriate permissions would gain access to the necessary resources. In the *TOMS* initiative, authors may provide access to remote systems they supply for the conduct of the review. In this sense, private resources are no longer an exclusion, but open access is also not an assumption.

Initiatives with mandatory reproductions also face the challenge of assessing reproducibility of research that relies on large-scale computational resources. One editor reflected on a recent case:

> We had another case where the author was very explicit that his computations take on the order of 20,000 compute hours and we just skipped that one, saying the data is all available, because it was a pure simulation, but we just can't run that raw data generation. It wasn't a complete

> failure, because he was kind enough as part of the replication archive to provide the output from those simulations. And so everything, the post-analysis and the table generation, we tested that part, but we didn't test the actual data creation. (Interviewee 3-1)

Another approach suggested by the *IS* initiative is for authors to provide detailed provenance information to demonstrate that the provided artifacts were used in the generation of reported results. Although not part of the reproducibility review process, the initiative suggests using automated provenance capture tools such as ReproZip to provide this information. Initiatives may consider obtaining access to computational resources through organizations such as XSEDE (Extreme Science and Engineering Discovery Environment) (Towns et al., 2014).

5.2. Advancing Reproducibility Review Through Software Tool Use and Development

Throughout our study of the seven reproducibility initiatives we have noted gaps in infrastructure, be it editorial management software, reproducibility and provenance tools, or reproduction frameworks. These gaps will remain for the foreseeable future and will be faced by new initiatives. The National Academies report Recommendation 6-3 exhorts funding agencies to invest in the "development of open-source, usable tools and infrastructure that support reproducibility for a broad range of studies across different domains in a seamless fashion." In this section, we consider the implications of our findings as they relate to funding new infrastructure development. We identify three distinct areas where infrastructure improvement is needed: 1) support for reproducibility review in editorial workflows, 2) reproducibility platforms, and 3) standards for artifact packaging and dissemination. These three categories can work in tandem to improve the review process significantly.

The studied initiatives demonstrate that the reproducibility review process is not well-suited for current editorial management tools. Reproducibility initiatives are developing infrastructure that may prove reusable, such as the AEA Jira workflow (Vilhuber et al., 2020) or the Confirmable Reproducible Research (CoRe2) project[13] underway at Odum. In the meantime, new initiatives will need to adopt new tools or adapt editorial workflows to handle operations, including reviewer assignment, progress tracking, and communicating reproducibility reports. While these capabilities may eventually be integrated into commercial editorial management platforms, there may be benefits to the creation of open-source tools that can be closely integrated with reproducibility tools and repository platforms.

Recent advancements in the development of infrastructure specifically to support computational reproducibility will likely play a role in ongoing and future initiatives. Platforms such as Binder (Jupyter et al., 2018), Code Ocean, ReproServer (Rampin et al., 2018), and Whole Tale (Chard, Gaffney, Jones, Kowalik, Ludäscher, Nabrzyski, et al., 2019) may be used to simplify and even at some point automate parts of the review and verification process. However, these tools and platforms will need to interoperate with existing peer review/editorial and repository infrastructure.

Funding agencies have invested significantly in the development of scientific workflow systems, automated provenance capture tools, virtualization technologies, as well as general-purpose reproducibility platforms (Brinckman et al., 2019). Evaluating these tools in the context of other initiatives may identify ways to improve them to achieve wider adoption. Research repositories have been demonstrated to be effective for the dissemination and preservation of computational artifacts; however, they do not directly support reproduction efforts. That research repositories feature centrally in the studied initiatives is a testament to their maturity, and some repositories are better suited for research from a specific domain or include features specific to integration with journals or publisher platforms. Repositories are rapidly evolving from their data-centric roots to better support publishing research codes, but preservation of the computational environment is today a limitation. Technologies exist to preserve information about the computational environment in binary form, such as virtualization or container technologies. However, due to their size, these images are generally unwelcome in research repositories. Public image registries exist today (e.g., Docker Hub), but do not provide the archival assurances of research infrastructure. Even if research repositories accepted these images, external infrastructure is required to support their creation and reexecution (e.g., Binder [Jupyter et al., 2018], ReproServer [Rampin et al., 2018], or Whole Tale [Chard, Gaffney, Jones, Kowalik, Ludäscher, Nabrzyski, et al., 2019]). This suggests an opportunity to better align reproducibility tools with the editorial infrastructure required for review, repository infrastructure required for dissemination, and computational infrastructure required for reexecution.

The integration of reproducibility tools into author, editorial, and publishing workflows highlights the need for relevant dissemination standards. Figure 1 illustrates the central role of reproducible computational research artifacts in the evolving scholarly publication process. Researchers must create these artifacts in conformance with initiative policies, often using tools created by research infrastructure developers. These artifacts become part of the scholarly record through research repositories or publisher platforms, which provide discovery capabilities. Journal editors and conference organizers establish the criteria and workflows that determine whether the provided artifacts are reproducible. Reviewers or verifiers certify reproducibility and, in some cases, assign badges or other metadata to artifacts to indicate that they have undergone additional assessment. Today, many reproducibility tools have defined their own formats for publishing reproducible research. For example, “binders” (Jupyter et al., 2018), “tales” (Chard, Gaffney, Jones, Kowalik, Ludäscher, Nabrzyski, et al., 2019), “sciunits” (That et al., 2017), “reprozips” (Chirigati, Rampin, et al., 2016), “capsules” (Code Ocean, 2020) just to name a few. The packages produced by these tools are often deposited into research repositories as common zip archives. While this practice ensures that the deposited package can easily be acted upon by the tool (e.g., reexecuted), it conceals key information from the repository that can be used for discovery (e.g., individual code and data files) and relies on external processes to assign relevant metadata (e.g., badges or certifications). Nascent open metadata standards, such as the Research Object Crate (RO-Crate) (Sefton et al., 2019) and the Whole Tale “Tale” (Chard, Gaffney, Jones, Kowalik, Ludäscher, McPhillips, et al., 2019) present an opportunity to define an information standard that supports the representation of these compound research objects (e.g., code, data, workflow, environment) that can be ingested into research repositories, but also the requirements of different actors in the assessment process (e.g., assignment of badges/certifications). To this end, we propose the concept of *assessable computational research artifacts* that contain all information required to perform a computational reproduction but also support verification and review metadata indicating how the artifact has been assessed. Instead of the ‘badge’ as an indicator on a paper or the metadata record stored in a research repository, it becomes an integral part of the object and travels with it. Assessable computational research artifacts therefore can stand alone while providing sufficient descriptive information to be understood, reproduced, and related to externally published resources.

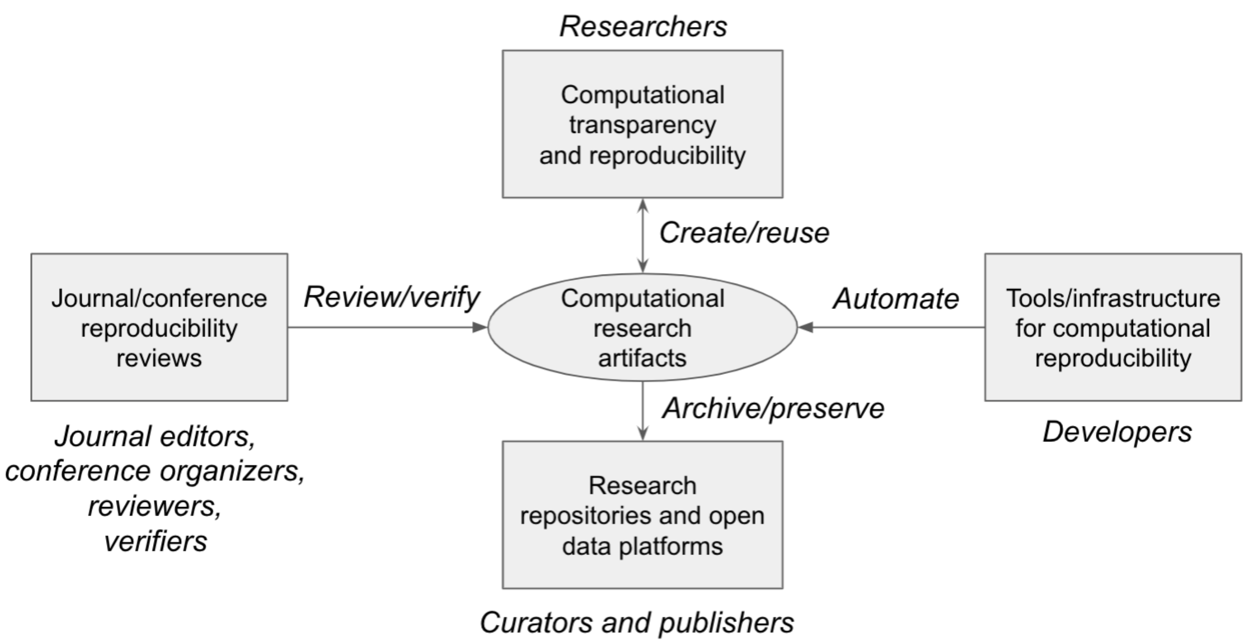


Figure 1. Assessable reproducible computational research artifacts.

Finally, improved instrumentation of the review process will not only aid in streamlining artifact assessment but also enable the measurement of initiative policy effects. By instrumenting the review process and even publishing the anonymized data that result, it may be possible for future researchers to study the broader effects of policies on the research and publication process. The NASEM recommendations discussed in this work and the studied initiatives suggest that improving the quality, rigor, and trustworthiness of results is best achieved by expanding the peer review process, which creates a significant burden on authors, editors, and reviewers. While these initiatives certainly increase the availability of computationally reproducible research artifacts, it does not necessarily follow that the quality or rigor of research is increased. In a critique of computational reproducibility policies, Drummond (2018) argues that, instead of increasing the burden on authors and reviewers, we should be increasing trust in reviewers and reducing their workloads: "[C]areful reviewing by experts is a much better defense against scientific misconduct than any execution of code" (p. 6). Leek and Peng (2015) argue that computational reproducibility is insufficient to address problematic research and instead argue for a "prevention" approach through increased education. Resnik & Shamoo (2017) argue that reproducibility is in part an ethical problem and the responsibility of the researcher, not necessarily the journal. Data provided by these and future initiatives can be used to study any broader effects of these policy changes.

6. Conclusions

The studied initiatives demonstrate how expanding the peer review process can be used to improve or even ensure the reproducibility of computational research at the time of publication. Mandatory full reproductions ensure that materials provided by authors can be used to reproduce reported results, but they come with a high cost today. Journal policies provide critical incentives to authors, but the verification process appears to be necessary to ensure policy compliance and the completeness of materials—hence "trust, but verify" in the title of this article. In this sense, the initiatives are consistent with earlier findings that, under current incentive structures, authors will not voluntarily provide these materials, and if they do, there are likely to be undetected ambiguities, errors, and oversights (Anderson & Dewald, 1994; Chang & Li, 2015; Dewald et al., 1986; Feigenbaum & Levy, 1993; Mirowski & Sklivas, 1991). If the goal is to ensure that materials provided by authors can be used to reproduce reported findings, then mandatory full reproductions provide the most comprehensive solution, assuming appropriate community readiness and an initiative with the resources to make this happen.

Whether these initiatives actually improve research quality and trustworthiness is an open question and opportunity for future work. Since the review process occurs post-acceptance, the initiatives may have limited impact on researcher practices. As noted by Leek and Peng (2015), reproducibility assessment at the point of publication is likely too late in the research process to affect upstream behaviors. Or perhaps these policies are part of a broader process of establishing discipline norms that, over time, will be further reflected in researcher practices. Nosek et al. (2012) express skepticism that such an extensive expansion of the peer review process is the best long-term general solution for improving the quality of computational research. Perhaps simple checklists may prove equally effective. Journals and conferences looking to adopt this approach should consider ways to measure the potential impact of the initiative on the overall quality of published research. Today, an easy way is to instrument the review process and expose the resulting data for analysis, opening the black box of peer review to future researchers.

The original *JMCB* study was conducted as an experiment on how changes in journal policy impact the availability and quality of research materials (Dewald et al., 1986). While these initiatives likely do affect the quality of research materials provided by authors, the question remains open as to whether they actually result in desirable effects on researcher behavior and improve the overall quality of published research. If it were possible to identify the policy used for the review of a particular paper, along with the number and types of errors identified during the review process, future work could potentially assess, by looking at citation and replication rates, whether policy changes have had desired effects. Communities considering implementing similar initiatives should consider not only internal operational metrics but also metrics that can be used to assess the overall impact of these types of efforts.

In this work we have presented the results of an investigation of seven reproducibility initiatives to better understand the steps that any new initiative would need to take in response to the recommendations of the 2019 National Academies reproducibility report. We developed a set of concrete decision points that can be used for new initiatives, identified key gaps in technical infrastructure, and pointed the way to an improved understanding of how changes to the incentives and information requirements of authors impact the quality, rigor, and trustworthiness of published computational research. Our findings clarify many of the "technological and practical challenges" suggested in NASEM Recommendation 6-4 while also highlighting the need for further study to better understand the impacts of these initiatives on the research and publication process. Our findings also speak to Recommendation 6-3 concerning investment in reproducibility infrastructure as well as Recommendation 6-5 concerning the dissemination of transparent research artifacts using research repositories (see Appendix E for the full text of the report recommendations). Without further expansion of editorial and repository infrastructure to better support the assessment and dissemination of computational research artifacts, new initiatives will continue to face significant technical obstacles in addition to the social challenges of expanding peer review requirements.

These initiatives—particularly the mandatory ones—no doubt increase the availability and quality of materials provided by authors, but whether they result in improved research quality, rigor, or trustworthiness ultimately remains an open research question.

7. Postscript: Reproducibility, Replicability, and Qualitative Research

There is an apparent irony in conducting a qualitative investigation on the topic of computational reproducibility. Qualitative research is not inherently computational and relies on interpretive methods that necessarily cannot guarantee that the same data and same methods can be used to draw exactly the same results. However, we strongly believe that the questions approached in this investigation were best suited to qualitative analysis. In qualitative research traditions, the focus has recently turned instead to research transparency (Elman et al., 2018; Elman & Kapiszewski, 2014) and emerging examples of verification of published qualitative research (Leighley, 2019). Returning to the NASEM definitions, qualitative conclusions fall under `replicability'—"obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." We have provided access to the complete set of interview instruments, codebooks, and descriptions of the analytical process, so that another researcher has the information needed to replicate this study, the highest standard of transparency applicable to qualitative research.

Disclosure Statement

The authors have nothing to disclose.

Acknowledgments

We thank all the interview participants whose careful and detailed observations and insights are central to the reported findings. This material is based upon work supported by National Science Foundation Award #1941443.

References

Alvarez, R. M., Key, E. M., & Núñez, L. (2018). Research replication: Practical considerations. *PS: Political Science & Politics*, *51*(2), 422–426. https://doi.org/10.1017/S1049096517002566

Anderson, R. G., & Dewald, W. G. (1994). Replication and scientific standards in applied economics a decade after the Journal of Money, Credit and Banking Project. *Federal Reserve Bank of St. Louis Review*, 67(6), 79–83. https://doi.org/10.20955/r.76.79-83

Anderson, R. G., Greene, W. H., McCullough, B. D., & Vinod, H. D. (2008). The role of data/code archives in the future of economic research. *Journal of Economic Methodology*, *15*(1), 99–119. https://doi.org/10.1080/13501780801915574

Ashenfelter, O., Haveman, R. H., Riley, J. G., & Taylor, J. T. (1986). Editorial statement. *The American Economic Review*, *76*(4), v. https://www.jstor.org/stable/1806060

Baggerly, Keith A., & Berry, D. A. (2011, January 1). Reproducible research. *AMSTATNEWS*. https://magazine.amstat.org/blog/2011/01/01/scipolicyjan11/

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452. https://doi.org/10.1038/533452a

Barba, L. A. (2018). *Terminologies for reproducible research.* ArXiv. http://arxiv.org/abs/1802.03311

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Bernanke, B. S. (2004). Editorial statement. *The American Economic Review*, *94*(1), 404. https://www.jstor.org/stable/3592790

Bonnet, P., Manegold, S., Bjørling, M., Cao, W., Gonzalez, J., Granados, J., Hall, N., Idreos, S., Ivanova, M., Johnson, R., Koop, D., Kraska, T., Müller, R., Olteanu, D., Papotti, P., Reilly, C., Tsirogiannis, D., Yu, C., Freire, J., & Shasha, D. (2011). Repeatability and workability evaluation of SIGMOD 2011. *SIGMOD Record*, *40*(2), 45–48. https://doi.org/10.1145/2034863.2034873

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B. D., Nabrzyski, J., & others. (2019). Computing environments for reproducibility: Capturing the "Whole Tale." *Future Generation Computer Systems*, *94*, 854–867. https://doi.org/10.1016/j.future.2017.12.029

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and reproducible research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics* (pp. 55–81). Springer. https://doi.org/10.1007/978-1-4612-2544-7_5

Chang, A. C., & Li, P. (2015). *Is economics research replicable? Sixty published papers from thirteen journals say "usually not"* (No. 2015–083; Finance and Economics Discussion Series). Board of Governors of the Federal Reserve System. https://dx.doi.org/10.17016/FEDS.2015.083

Chang, A. C., & Li, P. (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, *107*(5), 60–64. https://doi.org/10.1257/aer.p20171034

Chard, K., Gaffney, N., Jones, M. B., Kowalik, K., Ludäscher, B., McPhillips, T., Nabrzyski, J., Stodden, V., Taylor, I., Thelen, T., Turk, M. J., & Willis, C. (2019). Application of BagIt-Serialized Research Object Bundles for packaging and re-execution of computational analyses. *2019 15th International Conference on EScience (EScience)* (pp. 514–521). IEEE. https://doi.org/10.1109/eScience.2019.00068

Chard, K., Gaffney, N., Jones, M. B., Kowalik, K., Ludäscher, B., Nabrzyski, J., Stodden, V., Taylor, I., Turk, M. J., & Willis, C. (2019). Implementing computational reproducibility in the Whole Tale environment. P-RECS'19: *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems* (pp. 17–22). ACM. https://doi.org/10.1145/3322790.3330594

Chirigati, F., Capone, R., Rampin, R., Freire, J., & Shasha, D. (2016). A collaborative approach to computational reproducibility. *Information Systems*, *59*, 95–97. https://doi.org/10.1016/j.is.2016.03.002

Chirigati, F., Rampin, R., Shasha, D., & Freire, J. (2016). ReproZip: Computational reproducibility with ease. *Proceedings of the 2016 International Conference on Management of Data*, 2085–2088. https://doi.org/10.1145/2882903.2899401

Christian, T.-M., Lafferty-Hess, S., Jacoby, W. G., & Carsey, T. (2018). Operationalizing the replication standard. *International Journal of Digital Curation, 13*(1), 114–124. https://doi.org/10.2218/ijdc.v13i1.555

Claerbout, J. F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992* (pp. 601–604). Society of Exploration Geophysicists. https://doi.org/10.1190/1.1822162

Code Ocean. (2020). *What is a compute capsule?* http://help.codeocean.com/en/articles/1204225-what-is-a-compute-capsule

Collberg, C., & Proebsting, T. A. (2016). Repeatability in computer systems research. *Communications of the ACM*, *59*(3), 62–69. https://doi.org/10.1145/2812803

Committee on Reproducibility and Replicability in Science, Committee on National Statistics, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education, Nuclear and Radiation Studies Board, Division on Earth and Life Studies, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Analytics, Division on Engineering and Physical Sciences, Committee on Science, Engineering, Medicine, and Public Policy, Board on Research Data and Information, Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press. https://doi.org/10.17226/25303

DA-RT. (2015). Data Access and Research Transparency (DA-RT): A Joint statement by political science journal editors. *Political Science Research and Methods*, 3(3), 421-421. https://doi.org/10.1017/psrm.2015.44

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, *76*(4), 587–603. https://www.jstor.org/stable/1806061

Diggle, P. J., & Zeger, S. L. (2009). Editorial. *Biostatistics*, *10*(3), 405–408. https://doi.org/10.1093/biostatistics/kxp014

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., & Stodden, V. (2009). Reproducible research in computational harmonic analysis. *Computing in Science Engineering*, *11*(1), 8–18. https://doi.org/10.1109/MCSE.2009.15

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

Drummond, C. (2018). Reproducible research: A minority opinion. *Journal of Experimental & Theoretical Artificial Intelligence*, *30*(1), 1–11. https://doi.org/10.1080/0952813X.2017.1413140

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press. https://mitpress.mit.edu/books/vast-machine

Elman, C., & Kapiszewski, D. (2014). Data access and research transparency in the qualitative tradition. *PS: Political Science & Politics*, *47*(1), 43–47. https://doi.org/10.1017/S1049096513001777

Elman, C., Kapiszewski, D., & Lupia, A. (2018). Transparent social inquiry: Implications for political science. *Annual Review of Political Science*, *21*(1), 29–47. https://doi.org/10.1146/annurev-polisci-091515-025429

Eubank, N. (2016). Lessons from a Decade of replications at the Quarterly Journal of Political Science. *PS: Political Science & Politics*, *49*(2), 273–276. https://doi.org/10.1017/S1049096516000196

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*(11), 2628–2631. https://doi.org/10.1073/pnas.1708272114

Feigenbaum, S., & Levy, D. M. (1993). The market for (ir)reproducible econometrics. *Social Epistemology*, *7*(3), 215–232. https://doi.org/10.1080/02691729308578695

Ferro, N., & Kelly, D. (2018). SIGIR Initiative to implement ACM artifact review and badging. *ACM SIGIR Forum*, *52*(1), 4–10. https://doi.org/10.1145/3274784.3274786

Freire, J., Fuhr, N., & Rauber, A. (2016). Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041). *Dagstuhl Reports*, *6*(1), 108–159. https://doi.org/10.4230/DagRep.6.1.108

Fuentes, M. (2016, July 1). Reproducible research in *JASA*. *AMSTATNEWS*. https://magazine.amstat.org/blog/2016/07/01/jasa-reproducible16/

Fursin, G., & Dubach, C. (2014). Community-driven reviewing and validation of publications. *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering*, Article 5, 1–4. https://doi.org/10.1145/2618137.2618142

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue Canadienne d'économique*, *40*(3), 715–733. https://doi.org/10.1111/j.1365-2966.2007.00428.x

Harrell, S. L., Nam, H. A., Larrea, V. G. V., Keville, K., & Kamalic, D. (2015). Student Cluster Competition: A Multi-Disciplinary Undergraduate HPC Educational Tool. *Proceedings of the Workshop on Education for High-Performance Computing*. ACM. https://doi.org/10.1145/2831425.2831428

Heroux, M. A. (2015). Editorial: ACM TOMS Replicated Computational Results Initiative. *ACM Transactions in Mathematical Software*, *41*(3), Article 13. https://doi.org/10.1145/2743015

Heroux, M. A. (2019). *Trust me. QED.* Sandia National Lab. https://www.osti.gov/servlets/purl/1544811

Hopkins, T. (2009). The collected algorithms of the ACM. *WIREs Computational Statistics*, *1*(3), 316–324. https://doi.org/10.1002/wics.40

Ioannidis, J. P. A. (2005). *Why most published research findings are false*. *PLoS Med, 2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jacoby, W., G., Lafferty-Hess, S., & Christian, T.-M. (2017, July 17). Should journals be responsible for reproducibility? *Inside Higher Ed*. https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility

Jupyter, P., Bussonnier, M., Forde, J., Freeman, J., Granger, B., Head, T., Holdgraf, C., Kelley, K., Nalvarte, G., Osheroff, A., Pacer, M., Panda, Y., Perez, F., Ragan-Kelley, B., & Willing, C. (2018). Binder 2.0—Reproducible, interactive, sharable environments for science at scale. In F. Akici, D. Lippa, D. Niederhut, & M. Pacer (Eds.), *Proceedings of the 17th Python in Science Conference* (pp. 113–120). SciPy. https://doi.org/10.25080/Majora-4af1f417-011

Keiding, N. (2010a). Reproducible research and the substantive context. *Biostatistics*, *11*(3), 376–378. https://doi.org/10.1093/biostatistics/kxq033

Keiding, N. (2010b). Reproducible research and the substantive context: Response to comments. *Biostatistics*, *11*(3), 395–396. https://doi.org/10.1093/biostatistics/kxq034

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. https://doi.org/10.2307/420301

Konkol, M., Nüst, D., & Goulier, L. (2020). Publishing computational research—A review of infrastructures for reproducible and transparent scholarly communication. *Research Integrity and Peer Review*, *5*(1), Article 10. https://doi.org/10.1186/s41073-020-00095-y

Kovacevic, J. (2007). How to encourage and publish reproducible research. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, *4*, IV-1273-IV–1276. IEEE. https://doi.org/10.1109/ICASSP.2007.367309

Krafczyk, M., Shi, A., Bhaskar, A., Marinov, D., & Stodden, V. (2019). Scientific tests and continuous integration strategies to enhance reproducibility in the scientific software context. *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, 23–28. https://doi.org/10.1145/3322790.3330595

Krishnamurthi, S. (2013). Artifact evaluation for software conferences. *SIGSOFT Software Engineering Notes*, *38*(3), 7–10. https://doi.org/10.1145/2464526.2464530

Krishnamurthi, S., & Vitek, J. (2015). The real software crisis: Repeatability as a core value. *Communications of the ACM*, *58*(3), 34–36. https://doi.org/10.1145/2658987

Leek, J. T., & Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, *112*(6), 1645–1646. https://doi.org/10.1073/pnas.1421412111

Leighley, J. (2019, May 22). Verification, verification. *American Journal of Political Science*. https://ajps.org/2019/05/22/verification-verification/

LeVeque, R. J. (2006). Wave propagation software, computational science, and reproducible research. In M. Sanz-Sole et al. (Eds.), *Proceedings of the International Congress of Mathematicians*. (p. 1227-1254). European Mathematical Society. https://faculty.washington.edu/rjl/pubs/icm06

Manegold, S., Manolescu, I., Afanasiev, L., Feng, J., Gou, G., Hadjieleftheriou, M., Harizopoulos, S., Kalnis, P., Karanasos, K., Laurent, D., Lupu, M., Onose, N., Re, C., Sans, V., Senellart, P., Wu, T., & Shasha, D. (2010). Repeatability & workability evaluation of SIGMOD 2009. *SIGMOD Record*, *38*(3), 40–43. https://doi.org/10.1145/1815933.1815944

Manolescu, I., Afanasiev, L., Arion, A., Dittrich, J., Manegold, S., Polyzotis, N., Schnaitter, K., Senellart, P., Zoupanos, S., & Shasha, D. (2008). The repeatability experiment of SIGMOD 2008. *SIGMOD Record*, *37*(1), 39–45. https://doi.org/10.1145/1374780.1374791

McCullough, B. D., & Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *The American Economic Review*, *93*(3), 873–892. https://www.jstor.org/stable/3132121

McPhillips, T., Willis, C., Gryk, M. R., Nuñez-Corrales, S., & Ludäscher, B. (2019). Reproducibility by other means: Transparent research objects. *2019 15th International Conference on EScience (EScience)*, 502–509. https://doi.org/10.1109/eScience.2019.00066

Meier, K. J. (1995). Replication: A view from the streets. *PS: Political Science & Politics*, *28*(3), 456–459.

Mirowski, P., & Sklivas, S. (1991). Why econometricians don't replicate (although they do reproduce). *Review of Political Economy*, *3*(2), 146–163. https://doi.org/10.1080/09538259100000040

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Peng, R. D. (2009). Reproducible research and biostatistics. *Biostatistics*, *10*(3), 405–408. https://doi.org/10.1093/biostatistics/kxp014

Peng, R. D. (2010). Discussion of Keiding. *Biostatistics*, *11*(3), 393–394. https://doi.org/10.1093/biostatistics/kxq032

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Peng, R. D., Dominici, F., & Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology*, *163*(9), 783–789. https://doi.org/10.1093/aje/kwj093

Sefton, P., Carragáin, E. Ó., Soiland-Reyes, S., Corcho, O., Garijo, D., Palma, R., Coppens, F., Goble, C., Fernández, J. M., Chard, K., Gomez-Perez, J. M., Crusoe, M. R., Eguinoa, I., Juty, N., Holmes, K., Clark, J. A., Capella-Gutierrez, S., Gray, A. J. G., Owen, S., Williams, A. R., ... Thelen, T. (2019). *RO-Crate Metadata Specification 1.0* (Version 1.0.0). Zenodo. https://doi.org/10.5281/zenodo.3541888

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, *11*, Article 76. https://doi.org/10.3389/fninf.2017.00076

Radder, H. (1996). *In and about the world: Philosophical studies of science and technology*. SUNY Press.

Rampin, R., Chirigati, F., Steeves, V., & Freire, J. (2018). *Reproserver: Making reproducibility easier and less intensive*. ArXiv. https://arxiv.org/abs/1808.01406

Resnik, D. B., & Shamoo, A. E. (2017). Reproducibility and research integrity. *Accountability in Research*, *24*(2), 116–123. https://doi.org/10.1080/08989621.2016.1257387

Schreier, M. (2012). *Qualitative content analysis in practice* (Davis Library). Sage Publications. https://catalog.lib.unc.edu/catalog/UNCb7335714

Spiegelhalter, D. (2017). Trust in numbers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 948–965. https://doi.org/10.1111/rssa.12302

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*. https://doi.org/10.1287/isre.7.1.111

Stodden, V., Krafczyk, M. S., & Bhaskar, A. (2018). Enabling the verification of computational results: An empirical evaluation of computational reproducibility. *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems*, Article 3, 1–5. https://doi.org/10.1145/3214239.3214242

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

That, D. H. T., Fils, G., Yuan, Z., & Malik, T. (2017). *Sciunits: Reusable research objects*. ArXiv. http://arxiv.org/abs/1707.05731

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R., & Wilkins-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering*, *16*(5), 62–74. https://doi.org/10.1109/MCSE.2014.80

Vilhuber, L. (2018). *Reproducibility and Replicability in Economics*. Commissioned Paper. National Academies of Sciences, Engineering, and Medicine. https://www.nap.edu/resource/25303/Reproducibility% 20in%20Economics.pdf

Vilhuber, L. (2019). Report by the AEA data editor. *AEA Papers and Proceedings*, *109*, 718–729. https://doi.org/10.1257/pandp.109.718

Vilhuber, L., Turrito, J., & Welch, K. (2020). Report by the AEA data editor. *AEA Papers and Proceedings*, *110*, 764–775. https://doi.org/10.1257/pandp.110.764

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Willis, C. (2020a). *Trust, but verify: An investigation of methods of verification and dissemination of computational research artifacts for transparency and reproducibility* (Ph.D. thesis). University of Illinois at Urbana-Champaign.

Willis, C. (2020b). *Trust, but verify: An investigation of methods of verification and dissemination of computational research artifacts for transparency and reproducibility*. Harvard Dataverse. https://doi.org/10.7910/DVN/CKOGZM

Wilson, R. (2012). Note from the editor. *American Journal of Political Science*, *56*(3), 519. https://www.jstor.org/stable/23316003

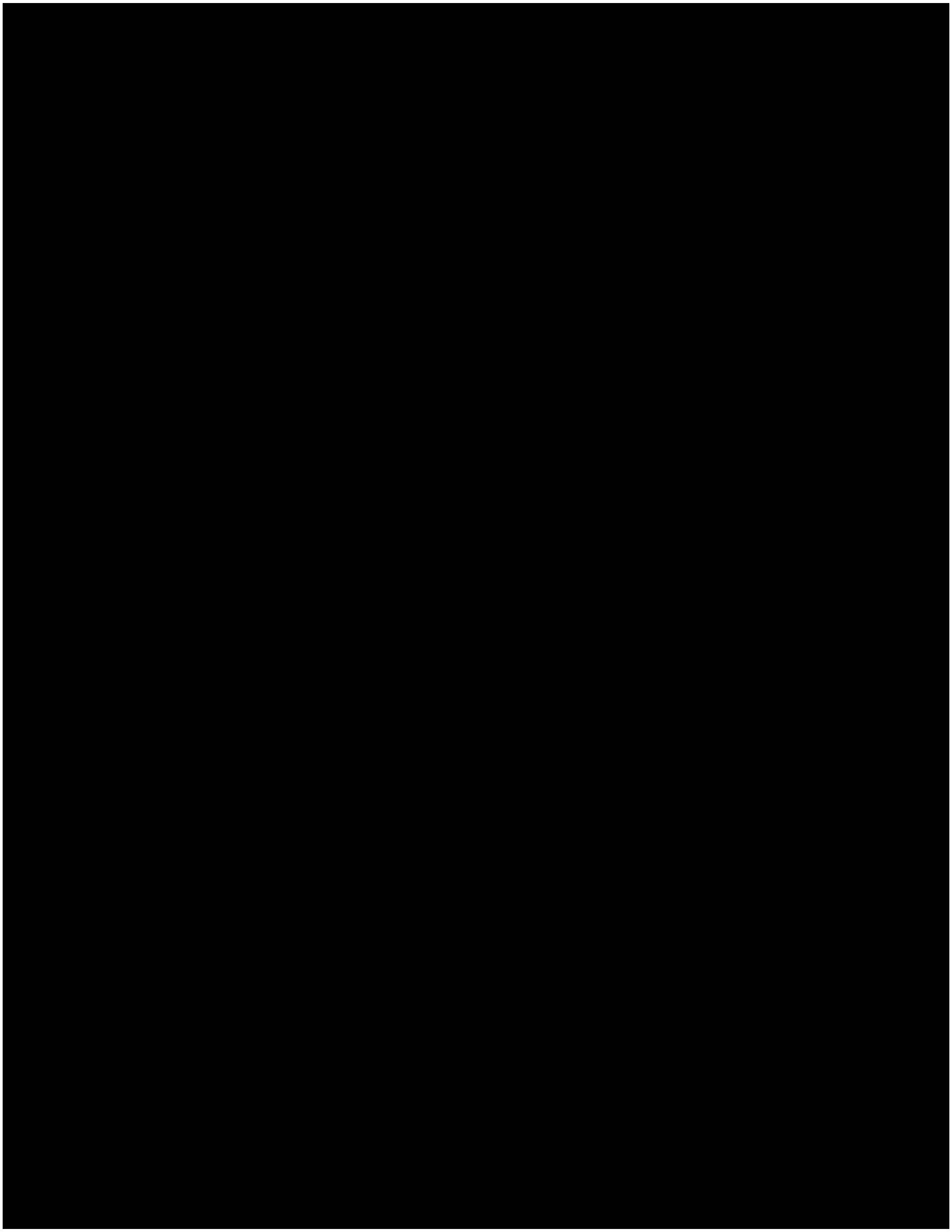
Yin, R. K. (2017). *Case study research and applications: Design and methods*. Sage Publications.

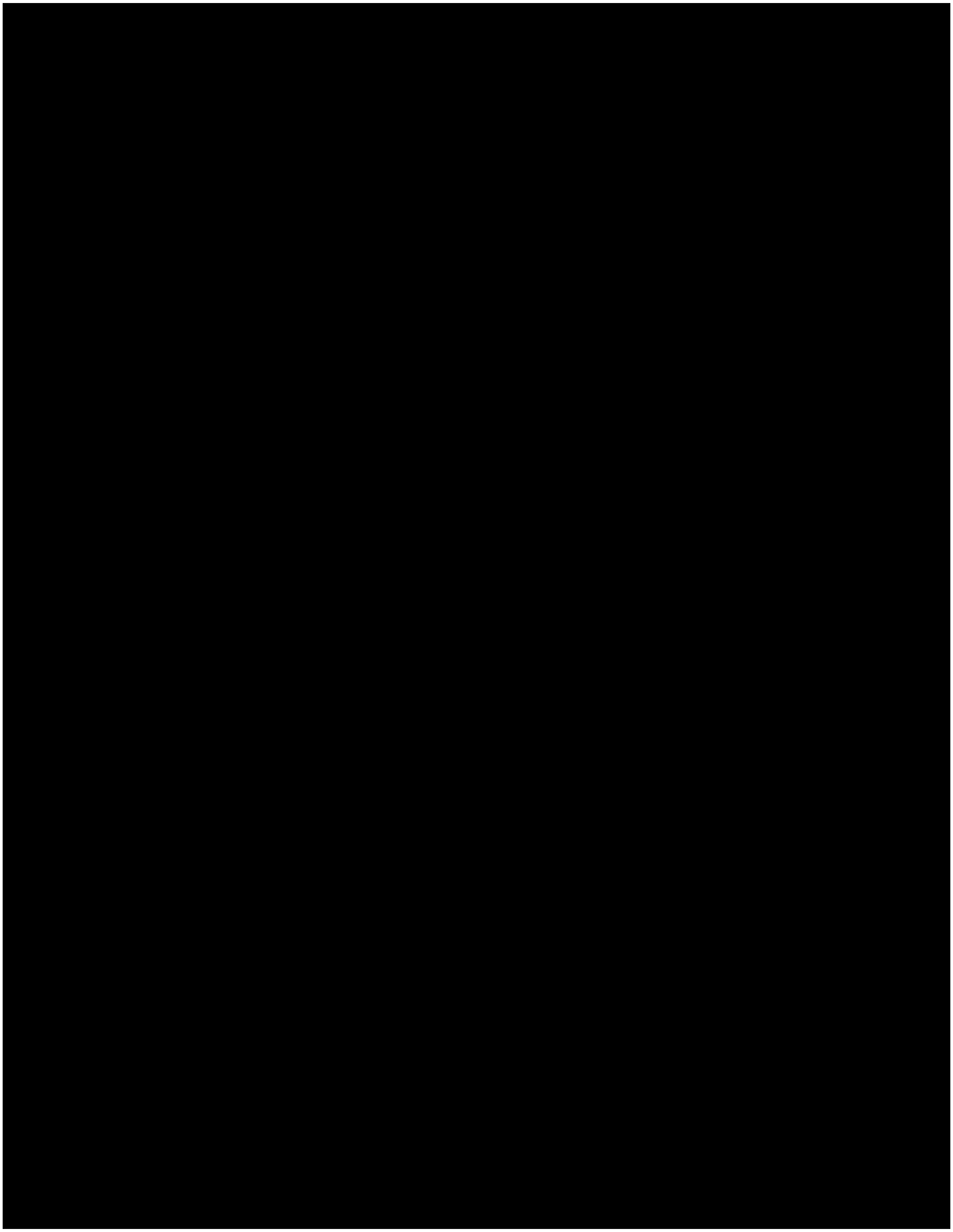
Yong, E. (2012, October 3). Nobel laureate challenges psychologists to clean up their act. *Nature*. https://doi.org/10.1038/nature.2012.11535

Appendix A

Interview Protocol

[redacted]





Appendix B

Documentary Evidence

Table B1 includes a complete listing of the documentary evidence used in the qualitative analysis.

Table B1. Documentary Evidence Used in Qualitative Coding.

Initiative	Document (Source)
AJPS	*AJPS* Verification Policy (https://ajps.org/ajps-verification-policy/) Replication and Verification Policy (https://ajps.org/wp-content/uploads/2019/03/ajps-replic-and-verif-policy-2-27-18.pdf) Guidelines for Preparing Replication Files (https://ajps.org/wp-content/uploads/2018/05/ajps_replication-guidelines-2-1.pdf) *AJPS* Dataverse (https://dataverse.harvard.edu/dataverse/ajps) Quantitative Data Verification Checklist (https://ajps.org/wp-content/uploads/2019/01/ajps-quant-data-checklist-ver-1-2.pdf) Qualitative Data Verification Checklist (https://ajps.org/wp-content/uploads/2019/01/ajps-qualdata-checklist-ver-1-0.pdf) Job advertisement (via Email) Journals_CurationChecklist.docx (Odum shared filesystem) Journals_CurationProcedures_Current.txt (Odum shared filesystem) Journals_VerificationChecklist.docx (Odum shared filesystem) JournalVerifier_NDA.docx (Odum shared filesystem) VM_Instructions_Verifier.docx (Odum shared filesystem)

AJPS Email Templates Examples.docx (Odum shared filesystem) Data Access and Research Transparency (DA-RT) (DA-RT, 2015) Anti-DART Petition (https://dialogueondartdotorg.files.wordpress.com/2015/11/petition-from-concerned-scholars-nov-12-2015-complete.pdf) *AJPS* Editorial Reports 2012-2019 (https://ajps.org/editor-reports/) Should Journals Be Responsible for Reproducibility? (Jacoby et al., 2017) Verification Verification (https://ajps.org/2019/05/22/verification-verification/) Our Experience with the *AJPS* Transparency and Verification Process for Qualitative Research (https://ajps.org/2019/ 05/09/our-experience-with-the-ajps-transparency-and-verification-process-for-qualitative-research) Celebrating Verification, Replication, and Qualitative Research Methods at the *AJPS* (https://ajps.org/2019/03/20/celebrating-verification-replication-and-qualitative-research-methods-at-the-ajps) Some Details about New AJPS Submission Requirements (https://ajps.org/2018/08/10/new-ajps-submission-requirements/) QDR (Qualitative Data Repository) and the *AJPS* Replication Policy (https://ajps.org/2016/11/22/qdr-and-the-ajps-replication-policy/) AJPS to Award COS Open Practice Badges (https://ajps.org/2 016/05/10/ajps-to-award-cos-open-practice-badges)	
AEA	Data and Code Availability Policy (https://www.aeaweb.org/journals/policies/data-code/)

AEA Data and Code Repository (https://www.openicpsr.org/openicpsr/aea)

Guidance on how to deposit data at the AEA Data and Code Repository (https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea-guidance.html)

Data and Code Availability Policy: Frequently Asked Questions (https://www.aeaweb.org/journals/policies/data-code/faq)

Verification guidance (https://social-science-data-editors.github.io/guidance/Verification_guidance.html)

Example replication report (https://github.com/AEADataEditor/replication-template)

Training and Guidance for assessing replicability (https://github.com/labordynamicsinstitute/replicability-training)

Unofficial guidance on various topics by the AEA Data Editor (https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea.html)

Report by the AEA Data Editor (Vilhuber, 2019)

Updated AEA Data and Code Availability Policy (July 16, 2019) (https://www.aeaweb.org/news/member-announcements-july-16-2019)

Reproducibility and Replicability in Economics (https://www.nap.edu/resource/25303/Reproducibility%20in%20Economics.pdf)

Workflow (https://github.com/labordynamicsinstitute/replicability-training/blob/master/jira-workflow-training.md)

Job posting (https://studentjobs.seo.cornell.edu/jobpostings/view?id=63161)

JASA	Reviewer Guidelines (via Email) *JASA-ACS* GitHub organization (https://github.com/jasa-acs/) Reproducible Research in JASA (https://magazine.amstat.org/blog/2016/07/01/jasa-reproducible16/) JASA Editors Talk Reproducibility (https://www.amstat.org/ASA/Publications /Q-and-As/JASA-Editors-Talk-Reproducibility.aspx) Author Contributions Checklist form Author Instructions (https://amstat.tandfonline.com/action/authorSubmission?journalCode=uasa20&page=instructions)
IS	Invited Reproducibility Papers - Author Guidelines (http://fchirigati.com/files/is/GuidelinesAuthors.txt) Invited Reproducibility Papers - Reviewer Guidelines (http://fchirigati.com/files/is/GuidelinesReviewers.txt) Guide for Authors (https://www.elsevier.com/wps/find/journaldescription.cws_home/236?generatepdf=true) A collaborative approach to computational reproducibility (Chirigati, Capone et al., 2016) New article type verifies experimental reproducibility (https://www.elsevier.com/connect/new-article-type-verifies-experimental-reproducibility)

Biostatistics	Information for Authors (https://academic.oup.com/biostatistics/pages/General_Instructions) Reproducible research and biostatistics (Peng, 2009) Editorial (Diggle & Zeger, 2009) Reproducible research and the substantive context (Keiding, 2010a) Discussion of Keiding (Peng, 2010) Reproducible research and the substantive context: Response to comments (Keiding, 2010b)
TOMS	The *TOMS* Initiative and Policies for Replicated Computational Results (RCR) (https://toms.acm.org/replicated-computational-results.cfm) Editorial: ACM *TOMS* Replicated Computational Results Initiative (Heroux, 2015) RCR Reviewer Invitation (via Email)

Paper submissions (https://sc19.supercomputing.org/submit/paper-submissions/)

Email - Appendix Review Instructions.pdf

Reproducibility Challenge Track (https://github.com/SC-Tech-Program/SCreproducibility/blob/master/Reproducibility-Challenge.md)

Journal Special Issue Track (https://github.com/SC-Tech-Program/SCreproducibility/blob/master/Journal-Special-Issue.md)

SC Reproducibility Materials (https://github.com/SC-Tech-Program/SCreproducibility)

Student Cluster Competition (http://www.studentclustercompetition.us/)

Student cluster competition: a multi-disciplinary undergraduate HPC educational tool (Harrell et al., 2015)

Parallel Computing special issue (SC16) (https://www.sciencedirect.com/science/article/pii/S0167819117301643)

Special Issue on SC17 Reproducibility Initiative (https://www.sciencedirect.com/science/article/pii/S0167819118302734)

Special Issue on the SC18 Student Cluster Competition Reproducibility Initiative (https://www.sciencedirect.com/science/article/pii/S0167819119301632)

Note. AEA = American Economic Association; *JASA-ACS* =*Journal of the American Statistical Association-Applications and Case Studies*; *Biostatistics* journal; *IS* = *Information Systems* journal; *SC* = *Supercomputing* conference; *TOMS* = *Transactions on Mathematical Software*.

Appendix C

Artifacts

AEA

1. Bernanke, B. 2020. *Data and code for: "The new tools of monetary policy."* American Economic Association. https://doi.org/10.3886/E117206V1
2. Bach, L., Laurent, C., & Sodini, P. 2020. *Rich pickings? Risk, return, and skill in household wealth.* American Economic Association. https://doi.org/10.3886/E117466V3
3. Farboodi, M., & Veldkamp, L. (2020). *Data and code for: Long run growth of financial data technology*. American Economic Association. https://doi.org/10.3886/E114984V2
4. Elder, T., & Zhou, Y. 2020. *Analysis code for the Black-White gap in non-cognitive skills among elementary school children*. American Economic Association. https://doi.org/10.3886/E117301V1
5. Bhandari, A., Birinci, S., McGrattan, E. R., & See, K. 2020. *Data and code for: What do survey data tell us about US businesses.* American Economic Association. https://doi.org/10.3886/E117021V3

AJPS

1. Casas, A., , Denny, M. J., & Wilkerson, J. 2020. More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process. *American Journal of Political Science, 64*(1), 5–18. https://doi.org/10.1111/ajps.12472, Data: https://doi.org/10.7910/DVN/7ZVSYO
2. Brierley, S., Kramon, E., & Kwaku Ofosu, G. 2020. The moderating effect of debates on political attitudes. *American Journal of Political Science, 64*(1), 19–37. https://doi.org/10.1111/ajps.12458, Data: https://doi.org/10.7910/DVN/OJA7YS
3. Haynes, K., & Yoder, B. K. 2020. Offsetting uncertainty: Reassurance with two-sided incomplete information. *American Journal of Political Science, 64*(1), 38–51. https://doi.org/10.1111/ajps.12464, Data: https://doi.org/10.7910/DVN/PXOT5L
4. Nielsen, R. A. 2020. Women's authority in patriarchal social movements: The case of female Salafi preachers. *American Journal of Political Science, 64*(1), 52–66. https://doi.org/10.1111/ajps.12459, Data: https://doi.org/10.7910/DVN/6YNZTE
5. Strickland, J. M. 2020. The declining value of revolving-door lobbyists: Evidence from the American states. *American Journal of Political Science, 64*(1), 67–81. https://doi.org/doi:10.1111/ajps.12485, Data: https://doi.org/10.7910/DVN/YQYZ6O

Biostatistics

Materials for these articles were no longer accessible at time of publication.

1. Lee, D., Ferguson, C., & Mitchell, R. 2009. Air pollution and health in Scotland: A multicity study. *Biostatistics, 10*(3), 409–423. https://doi.org/10.1093/biostatistics/kxp010
2. Magi, A., Benelli, M., Marseglia, G., Nannetti, G., Scordo M.R., Torricelli, F. 2010. A shifting level model algorithm that identifies aberrations in array-CGH Data. *Biostatistics, 11*(2), 265–280.

https://doi.org/10.1093/biostatistics/kxp051

3. Riebler, A., & Held, L. 2010. The analysis of heterogeneous time trends in multivariate age-period-cohort models. *Biostatistics, 11*(1), 57–69. https://doi.org/10.1093/biostatistics/kxp037
4. Varin, C., & Czado, C. 2010. A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics, 11*(1), 127–138. https://doi.org/10.1093/biostatistics/kxp042

Information Systems

1. Wolke, A., Bichler, M., Chirigati, F., & Steeves, V. 2016. Reproducible experiments on dynamic resource allocation in cloud data centers. *Information Systems, 59,* 98–101. https://doi.org/10.1016/j.is.2015.12.004
2. Lastra-Diaz, J. J., García-Serrano, A., Batet, M., Fernández, M., & Chirigati, F. 2017. HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems, 66,* 97–118. https://doi.org/10.1016/j.is.2017.02.002
3. Fariña, A., Martínez-Prieto, M. A., Claude, F., Navarrod, G., Lastra-Díaz, J. J., Prezza, N., & Seco, D. 2019. On the reproducibility of experiments of indexing repetitive document collections. *Information Systems, 83,* 181–194. https://doi.org/10.1016/j.is.2019.03.007

JASA-ACS

1. Banerjee, T., Mukherjee, G., Dutta, S., & Ghosh, P. 2019. A large-scale constrained joint modeling approach for predicting user activity, engagement, and churn with application to freemium mobile games. *Journal of the American Statistical Association, 115*(530), 538–554. https://doi.org/10.1080/01621459.2019.1611584
2. Lee, C., & Wilkinson, D. J. 2020. A hierarchical model of nonhomogeneous Poisson processes for Twitter retweets. *Journal of the American Statistical Association, 115*(529), 1–15. https://doi.org/10.1080/01621459.2019.1585358
3. Smith, A. N., & Allenby, G. M. 2020. Demand models with random partitions. *Journal of the American Statistical Association, 115*(529), 47–65. https://doi.org/10.1080/01621459.2019.1604360
4. Tang, X., Yang, Y., Yu, H. J., Liao, Q.-H., & Bliznyuk, N. 2019. A spatio-temporal modeling framework for surveillance data of multiple infectious pathogens with small laboratory validation sets. *Journal of the American Statistical Association, 114*(528), 1561–1573. https://doi.org/10.1080/01621459.2019.1585250
5. Wilson, D. R., Jin, C., Ibrahim, J. G., & Sun, W. 2019. ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *Journal of the American Statistical Association, 115*(531), 1055–1065. https://doi.org/10.1080/01621459.2019.1654874

Supercomputing

1. Ben-Nun, T., de Fine Licht, J., Ziogas, A. N., Schneider, T., Hoefler, T. 2019. Stateful dataflow multigraphs: A data-centric model for performance portability on heterogeneous architectures. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '19*. Association for Computing Machinery. https://doi.org/10.1145/3295500.3356173
2. Domke, J. et al. 2019. HyperX topology: First at-scale implementation and comparison to the fat-tree. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '19*. Association for Computing Machinery. https://doi.org/10.1145/3295500.3356140
3. Laguna, I. et al. 2019. A large-scale study of MPI usage in open-source HPC applications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '19*. Association for Computing Machinery. https://doi.org/10.1145/3295500.3356176
4. Li, L., & Chapman, B. 2019. Compiler assisted hybrid implicit and explicit GPU memory management under Unified Address Space. *In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '19*. Association for Computing Machinery. https://doi.org/10.1145/3295500.3356141
5. Narra, K. G. et al. 2019. Slack squeeze coded computing for adaptive straggler mitigation. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '19*. Association for Computing Machinery. https://doi.org/10.1145/3295500.3356170

TOMS

1. Willenbring, J..N. (2015). Replicated Computational Results (RCR) report for "BLIS: A framework for rapidly instantiating BLAS functionality?" *ACM Transactions on Mathematical Software,* 41(3), Article 15 (pp. 1-4). https://doi.org/10.1145/2738033
2. Bavier, E. (2016). Replicated Computational Results (RCR) report for *A sparse symmetric indefinite direct solver for GPU architectures. ACM Transactions on Mathematical Software,* 42(1), Article 2 (pp. 1-10). https://doi.org/10.1145/2851489
3. Meiser, D., (2016). Replicated Computational Results (RCR) report for "A distributed-memory package for dense hierarchically semi-separable matrix computations using randomization." *ACM Transactions on Mathematical Software* , 42(4), Article 28 (pp. 1-5). https://doi.org/10.1145/2929907
4. Lindquist, N. (2019). Replicated Computational Results (RCR) report for "Code generation for generally mapped finite elements." *ACM Transactions on Mathematical Software* , 45(4), Article 42 (pp. 1-7). https://doi.org/10.1145/3360984

Appendix D

Codebooks

This appendix includes the case profile structure and high-level codebooks used for the qualitative analysis of interview transcripts and documentary evidence. Complete codebooks are available at Willis (2020a, 2020b).

Table D1. Case Profile Structure.

Profile Section	**Description**
Initiative organization	How the initiative is organized including relationships to parent and other stakeholder organizations such as funding bodies, publishers, archive, etc.
Historical antecedents	Review of historical developments leading to the creation of the initiative within the specific organization and discipline.
Policy and guidelines	Summary of policy and guideline materials.
Technical infrastructure	Summary of technical infrastructure used by the initiative.
Artifacts, identifiers, badges, metadata	Summary of artifacts produced as part of the initiative as well as how the initiative applies identifiers, badges, and additional metadata related to the review process.
Initiative metrics	Summary of metrics used to measure initiative effects

Table D2. High-Level Codes Groups Used for Coding of Interview Transcripts.

Code Group	**Description**
Benefits	Discussion of benefits of the initiative to stakeholders including authors, reviewers, verifiers, curators, as well as journals, funders, and the interviewee themselves.

Challenges	Discussion of challenges encountered during the initiative including awareness; burden on authors, editors, and reviewers; gaps in infrastructure; cost; impact on publication review time; as well as use of students.
Community Response	Discussion of how the research community and stakeholders have reacted to the initiative.
Definitions	Interviewee definitions of reproducibility, replicability, and transparency.
Expertise	Discussion of expertise requirements for authors, editors, reviewers, and verifiers.
Measurement	Discussion of metrics used or considered to assess the effectiveness or impact of the initiative. This includes journal metrics (e.g., impact factor, submission rates, publication times) as well as others (e.g., download rates, errors found during review, survey responses).
Motivations	Discussion of the underlying motivation of the initiative.
Of What	Discussion of *what* is being reproduced or assessed for reproducibility in the defined workflow.

Table D3. High-Level Qualitative Codebook Categories Developed for Coding of Policies, Guidelines, and Checklists.

Code Group	**Description**
Reproducibility	Guidelines related to the reproduction or reproducibility assessment process including reviewer expertise, modes of reproduction, suitability, and access to resources.
Documentation	Guidelines related to general documentation such as README files, manifests, and computational workflows.

Software	Guidelines related to author-supplied software including accessibility, persistence, licenses, versions, documentation, and exceptions (e.g., proprietary source code).
Data	Guidelines related to source and analysis data, including accessibility, persistence, licenses, versions, documentation, formats, variable labeling, and exceptions (e.g., protected or proprietary source code).
Environment	Guidelines related to specification of the environment including accessibility (including external systems), software dependencies, operating system, hardware dependencies, compilers, runtime conditions, resource requirements, and exceptions (e.g., protected or proprietary source code).
Experimental context	Guidelines related to documentation of experiments including workflows/protocols, evaluation procedures, metrics, parameters (including random seed values), as well as robustness (e.g., experiment customization).
Results	Guidelines related to the accessibility and documentation of results including provenance information
Publication	Guidelines related to publishing artifacts including packaging, distribution, use of persistence identifiers, use of archival formats.

Appendix E

Selected NASEM Recommendations

This appendix includes the full-text of the National Academies recommendations referenced in this article. For further details see (Committee on Reproducibility and Replicability in Science et al., 2019).

RECOMMENDATION 6-3

> Funding agencies and organizations should consider investing in research and development of open-source, usable tools and infrastructure that support reproducibility for a broad range of

> studies across different domains in a seamless fashion. Concurrently, investments would be helpful in outreach to inform and train researchers on best practices and how to use these tools.

RECOMMENDATION 6-4

> Journals should consider ways to ensure computational reproducibility for publications that make claims based on computations, to the extent ethically and legally possible. Although ensuring such reproducibility prior to publication presents technological and practical challenges for researchers and journals, new tools might make this goal more realistic. Journals should make every reasonable effort to use these tools, make clear and enforce their transparency requirements, and increase the reproducibility of their published articles.

RECOMMENDATION 6-5

In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the National Science Foundation (NSF) should:

- Develop a set of criteria for trusted open repositories to be used > by the scientific community for objects of the scholarly record.
- Seek to harmonize with other funding agencies the repository > criteria and data-management plans for scholarly objects.
- Endorse or consider creating code and data repositories for > long-term archiving and preservation of digital artifacts that > support claims made in the scholarly record based on NSF-funded > research. These archives could be based at the institutional level > or be part of, and harmonized with, the NSF-funded Public Access > Repository.
- Consider extending NSF's current data-management plan to include > other digital artifacts, such as software.
- Work with communities reliant on non-public data or code to develop > alternative mechanisms for demonstrating reproducibility

Through these repository criteria, NSF would enable discoverability and standards for digital scholarly objects and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

Footnotes

1. The report states that *reproducibility* is the equivalent of computational reproducibility, so we define transparency as computational transparency. ↩
2. While the AEA-wide policy was implemented in 2019, previous journal-level policies had been established for many years. ↩
3. The *SC* initiative includes two distinct subinitiatives, the Artifact Description/Artifact Evaluation (AD/AE, https://sc19.supercomputing.org/submit/reproducibility-initiative/) and the Student Cluster Competition Reproducibility Challenge (SCC RC) (Harrell et al., 2015) . The AD/AE review is the primary focus for this study. ↩
4. Informants were selected because of their role in defining or implementing initiative policies and workflows. As such, authors were excluded. While they participate in these initiatives, they are not directly involved in their operationalization. ↩
5. The AEA and *AJPS* initiatives are examples of initiatives that have a very long history, as journal policies have evolved and become even stricter over many years. The *SC* initiative is an example of an initiative that adopted mandatory assessment after several years of an opt-in policy. ↩
6. Both the *IS* and *TOMS* initiatives consider the reproducibility paper or RCR report as reviewer incentives. By participating in the reproducibility review, reviewers gain a publication in the journal. ↩
7. ReproZip was developed for use by the database community by leaders in the *IS* initiative (Chirigati, Rampin et al., 2016). ↩
8. Resubmissions reflect the number of times the artifacts have been resubmitted for review and are a proxy for errors. ↩
9. Eubank (2016) reported a cost of US$180 for a single paper in a similar initiative. ↩
10. Jira is a commercial software project tracking platform (https://www.atlassian.com/software/jira). ↩
11. This is, in fact, what is already happening with the Artifact Description appendix within the ACM/IEEE community. The AD/AE appendix developed as part of the https://ctuning.org/ initiative serves as the basis for the *JASA-ACS* and *SC* initiatives. ↩
12. See https://github.com/craig-willis/reproducibility-checklist/ ↩

13. See https://odum.unc.edu/2018/07/alfred-p-sloan-foundation-grant/ ↩