TRUST, BUT VERIFY: AN INVESTIGATION OF METHODS OF VERIFICATION AND DISSEMINATION OF COMPUTATIONAL RESEARCH ARTIFACTS FOR TRANSPARENCY AND REPRODUCIBILITY

BY

CRAIG WILLIS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Professor Victoria Stodden, Chair
Professor Peter Darch
Professor Bertram Ludäscher
Professor Michela Taufer, University of Tennessee, Knoxville

Abstract

In this study, I investigate how research communities are addressing concerns about the quality and rigor of computational research. I focus specifically on initiatives to expand the peer review and publication process to include new requirements for the assessment and dissemination of computational research artifacts. I report the results of a multiple-case analysis of two primary (*American Journal of Political Science*, ACM/IEEE *Supercomputing*) and five supplemental cases in political science, computer science, economics, mathematics, and statistics. Cases were developed through qualitative analysis based on interviews with key stakeholders ($n = 17$) including editors, reviewers, and verifiers; a sample of ($n = 27$) verified artifacts; and documentary evidence including policies, guidelines, and workflows.

The central argument of this dissertation is that these reproducibility initiatives represent a set of experiments across the sciences exploring how changes to the incentives and information requirements of authors impact the quality, rigor, reproducibility, and trustworthiness of published research. These initiatives are part of a broader effort to change community norms with respect to the dissemination of the results of research that involves computation, elevating the importance of computational artifacts and clearly signaling that authors cannot be trusted to provide this information voluntarily. Expanding peer review and increasing the information required of authors through publication reproducibility audits is just one approach – and a costly one – to improving research quality and trustworthiness. The effect of these changes on research quality has yet to be demonstrated or studied.

Based on the cases, I identify key factors that influence the operationalization of policies and workflows; the elements that each community considers important to the assessment of computational transparency and reproducibility; as well as the tools and infrastructure that they leverage to aid in the creation, assessment and dissemination of reproducible research artifacts. I develop a framework to analyze the reproducibility initiatives and a conceptual model of reproducible research artifacts. I relate my findings to recommendations from the recent National Academies of Science, Engineering and Medicine (NASEM) report on *Reproducibility and Replicability in Science* and provide a set of normative guidelines for communities interested in pursuing similar initiatives with implications for journal and conference leadership; tool and infrastructure developers; and funding bodies. I conclude that, while promising, further efforts should be made to increase our understanding of the effect of initiative policies and technological advancements on research quality.

Acknowledgements

This work would not have been possible without the support and guidance of my committee and participation of the interviewees. I owe an immense debt of gratitude to Drs. Victoria Stodden and Bertram Ludäscher for encouraging and supporting me in the pursuit of these questions and for sharing so much of their time and insight. Many thanks to committee members Drs. Peter Darch and Michela Taufer for sharing their extensive expertise in qualitative and high-performance computing research respectively as well as their admirable attention to detail. I also thank all of the interview participants whose observations and insights are central to the reported findings.

This study draws on my experience with the Whole Tale project, where I've benefited from the expertise (and skepticism) of many project collaborators including Drs. Ludäscher and Stodden as well as Kacper Kowalik, Timothy McPhillips, Michael Hategen, Matthew B. Jones, Tommy Thelen, Matthew Turk, and Kyle Chard.

I also thank the team at the Odum Institute for Research in Social Science for their cooperation and support; Kenton McHenry at the National Center for Supercomputing Applications (NCSA) for tolerating me in completing this work; and Dr. Amy Marshall-Colon for helping me think through the practical implications.

The study presented here builds on earlier work on the study of scientific metadata and comes after a lengthy foray into experimental information retrieval research, which perhaps seeded my interest in the problems of computational reproducibility. I would like to thank Drs. Miles Efron, Garrick Sherman, Jana Diesner, Stephen Downie, Diane Kelly, Jaime Arguello for their support and guidance during that prior life. I would also like to thank Drs. Robert M. Losee and Jane Greenberg for setting me on this very fulfilling course.

Finally, I am of course indebted to my wife, Cheryl, for her unwavering support in the face of so many challenges and changes including the final chapter written in the mountains in the midst of a pandemic. *Il faut cultiver notre jardin.*

For Cannon and Alma. In memory of my father.

Contents

List of Tables

List of Figures

List of Abbreviations

ACM	Association for Computing Machinery
AD	Artifact Description (appendix)
AE	Artifact Evaluation (appendix)
AEA	American Economics Association
AER	American Economic Review
AJPS	American Journal of Political Science
ASA	American Statistical Association
CACM	Communications of the ACM
CALGO	Collected Algorithms of the ACM
CISER	Cornell Institute for Social and Economic Research
COS	Center for Open Science
CRA	Computational Results Analysis (appendix)
HPC	High-performance computing
ICPSR	Inter-univerisity Consortium for Political and Social Research
IEEE	Institute of Electrical and Electronics Engineers
IS	Information Systems (journal)
JAE	Journal of Applied Econometrics
JASA	Journal of the American Statistical Association
JCR	Journal Citation Reports
JMCB	Journal of Money, Credit and Banking

JIF	Journal Impact Factor
NASEM	National Academics of Science, Engineering, and Medicine
NSF	National Science Foundation
OSF	Open Science Framework
QDR	Qualitative Data Repository
PSE	Problem Solving Environment
SC	Supercomputing
SCC	Student Cluster Competition
SIGHPC	Special Interest Group on High Performance Computing
SIGMOD	ACM Special Interest Group for the Management of Data
TOMS	ACM Transactions on Mathematical Software
VCS	Version control system

Chapter 1

Introduction

Доверяй, но проверяй (Trust, but verify) — Russian Proverb

Over the past decade, increasing concerns about the reproducibility and replicability of published scientific findings have culminated in claims of a "reproducibility crisis" potentially eroding trust in science [20, 81, 89, 134, 199]. Failures to reproduce key findings in high-profile studies, many in the social sciences [39, 59, 60, 121] and medicine [15, 28, 222], have led research communities across disciplines to consider new practices to improve research transparency and rigor. Recommendations have included improvements to study design, larger studies, and a decreased reliance on p-values in reporting statistical significance [134]; preregistration and preanalysis plans to reduce publication, reporting, and selection bias [60]; improved training in statistics and data handling[1]; and stricter requirements for data and code sharing in support of transparency, reproducibility, and replicability [13, 14]. For the many communities that leverage data-driven and computational methods, there has been an increase in the adoption of data and code sharing practices and policies [257]. In several cases, journals and conferences have established policies requiring the sharing of complete computational workflows subject to publication audits to confirm that they can be re-executed to reproduce reported findings [57, 98, 122, 135, 211, 267]. This practice of sharing data, code, and computational workflows for review and verification represents a major expansion of publication and peer review processes, increasing information requirements on authors and editor and reviewer workloads. This has led to the development of new tools,

[1] For example, `https://grants.nih.gov/policy/reproducibility/training.htm`

formats, and infrastructure to reduce burden while facilitating dissemination and archiving [38, 50, 56, 97, 147, 262].

In 2017, in response to the growing concerns about eroding public trust in science, Congress directed the National Science Foundation (NSF) to engage with the National Academies of Sciences, Engineering, and Medicine (NASEM) to "assess reproducibility and replicability in scientific and engineering research and to provide findings and recommendations for improving rigor and transparency" [199]. The resulting consensus report on "Reproducibility and Replicability in Science" proposes cross-discipline definitions for *reproducibility* and *replicability* and makes recommendations for improving both in practice. In the report, *reproducibility* is defined as "obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis," which the authors assert is equivalent to *computational reproducibility*; while *replicability* is "obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." The narrow focus on computational reproducibility is due, they say, to the increasing role of computation in science combined with a lack of uniformity in how scientists develop research software and share computational results. While computational reproducibility is the focus of the study presented here, it is important to recognize that reproducibility in this sense is often in the service of the broader notion of replicability.

The NASEM report also notes the strong relationship between transparency and reproducibility. Transparency, according to the report, "represents the extent to which researchers provide sufficient information to enable others to reproduce the results." In other words, transparency is required but not sufficient to guarantee reproducibility. This suggests that computational reproducibility is on a spectrum with transparency on one end (making code and data available) and re-executability on the other (making sure that computations can be re-executed to produce results). This distinction between reproducibility-as-transparency and reproducibility-as-re-executability has implications for publication reproducibility audits, tools, and packaging formats.

The NASEM report makes a series of recommendations for improving computational reproducibility in practice. Four of the recommendations are necessary for any research community interested in supporting the verification and dissemination of reproducible computational artifacts. These include[2] 1) requirements of authors to provide sufficient information to enable computational reproducibility [4-1] ; 2) further investment in the development of cross-domain tools and infrastructure in support of computational reproducibility [6-3]; 3) the implementation of publication reproducibility audits[3] [6-4]; and 4) the use of archival repositories and open data platforms for sharing, preservation and re-use of related artifacts [6-5]. These recommendations are informed by a number of existing discipline- and community-driven initiatives such as the development of tools to aid researchers in the creation and publication of computationally reproducible research artifacts [38, 52, 56, 262, 197, 101]; the adoption of policies and workflows for the review and verification of artifacts during the peer review process [55, 57, 98, 122, 211, 267] and standards for the representation of these scholarly artifacts for dissemination and long term archiving [53, 55, 196, 262]. Research communities interested in adopting these recommendations will benefit from specific guidance, which I provide in the final chapter of this study.

The NASEM definitions and recommendations add to a growing discussion about the precise meanings of the terms reproducibility and transparency across the sciences. These terms have long been used by different scientific communities and carry different and often contradicting meanings [21, 111, 123, 219]. With respect to reproducibility, by foregrounding computation the NASEM definition seemingly de-emphasizes reproducibility of the non-computational aspects of research [186]. This raises questions about the relationship between computational reproducibility to the broader concepts of replicability and scientific reproducibility in general. For many researchers who rely largely on computation in their work, the narrow concept of computational reproducibility has historically been well received

[2] The full text of these recommendations is provided in Appendix A.

[3] "Publication reproducibility audits" are formal review or assessment processes adopted by journals to ensure computational reproducibility.

[80, 145, 213]. However, for those who rely on computation only as part of a larger research process, the narrow focus on computation is seen as problematic [125, 141, 186]. In the field of computer science, the issue of computational reproducibility is perhaps even more existential – encompassing the reproducibility of research in the field as a whole [95]. The NASEM definition of reproducibility is still open to interpretation and operationalization. The interpretation is critical when considering the impact on the development and adoption of tools and infrastructure for computationally reproducible research, the operationalization of publication reproducibility audits, and the dissemination of reproducible research artifacts via archival repositories. A key problem when conceptualizing reproducibility in computational research is whether the computational environment is the instrument used to study other phenomena or whether the environment itself is the object of study.

The situation can be seen through the comparison of two different approaches to confirming computational reproducibility from the two cases that are the focus of the research presented here: the *American Journal of Political Science* (AJPS) and the ACM *Supercomputing* (SC) conference. As part of its ongoing reproducibility initiative, the Supercomputing conference assesses the reproducibility of submitted papers through the review of additional information provided in a mandatory supplemental appendix for all technical papers. By contrast, the *AJPS* requires authors to deposit all materials required to reproduce analytical results into a central archival repository subject to pre-publication verification through re-execution and comparison of results to figures, tables, and in-text claims in the accepted manuscript. Such a detailed reproduction is impractical under conference timelines and impossible for many SC papers, as they rely on large-scale computational resources, including boutique and leadership-class systems that are accessible to only a handful of researchers worldwide. While both of these operationalizations fall under a broad definition of computational reproducibility, they are quite different in their motives, constraints, implementations and outcomes. They also reflect the readiness of their respective communities regarding the social and technical infrastructure required.

The central argument of this dissertation is that the reproducibility initiatives that are the focus of this study are a set of parallel experiments across the sciences exploring how changes to the incentives and information requirements of authors impact the quality, rigor, and trustworthiness of published computational research. As with many activities related to scholarly communication, peer review, and publication, they are not seen as such nor are they designed with a clear hypothesis or means of measurement. While their motives are laudable, these initiatives place increased burden on researchers, editors, and reviewers – sometimes at a substantial cost. Communities interested in pursuing similar activities, in light of the absence of concrete evidence of their effectiveness, should be aware of current gaps in editorial and publication infrastructure that may impede their success. Developers of general-purpose tools and infrastructure for computational reproducibility may also benefit from an increased focus on communities with a high level of readiness. The results of this study suggest an opportunity to demonstrate the utility of these advancements with the initiatives that are best-suited to benefit from their capabilities.

1.1 Research Questions and Approach

This study takes as its starting point the four NASEM recommendations summarized above to further clarify the concepts of computational transparency and computational reproducibility as they relate to methods of verifying and disseminating computational research artifacts. This is achieved through a multiple case analysis of seven "reproducibility initiatives" designed to improve computational reproducibility across the fields of political science, computer science, economics, mathematics, and statistics. I define *reproducibility initiative* as formal activities undertaken primarily by journal editors, conference organizers, or related stakeholders to improve the transparency and reproducibility of research published through their venues. These initiatives come in many forms, but typically involve new requirements for authors to publish materials beyond the manuscript, as well as the creation of new roles, expanding the peer review process to include the review and assessment of these materials. I define *computational research artifacts*[4] as the packaged research artifacts that are generated and reviewed as a result of there processes. The study focuses on the operational workflows adopted by these communities with the ultimate goal of clarifying the structure of the verified and packaged artifacts as they relate to and are informed by each community's goals, operationalized publication audits, and related tools and infrastructure.

This study addresses three research questions:

- RQ1. How are computational transparency and computational reproducibility operationalized through publication reproducibility audits?
- RQ2. What are the characteristics of research artifacts that make them computationally reproducible (or irreproducible)?
- RQ3. What are the key characteristics of tools and packaging formats[5] that enable computational transparency and reproducibility?

These questions are considered in the context of the four NASEM recommendations discussed above. For authors and journals to ensure computational reproducibility (recommendations 4-1 and 6-4), we must understand what it means for published artifacts to be reproducible (RQ2) and how to operationalize review or assessment processes (RQ1). In order to develop tools and infrastructure to better support computational reproducibility (recommendation 6-3) we need to better understand the requirements of researchers and journals. To facilitate sharing of computational research artifacts (recommendation 6-5) we myst clarify information is required and how it can be represented for archiving and dissemination (RQ3).

[4] I've found the concept of *research objects* in the broad sense described in [24] to be quite intuitive for these purposes ("a class of artefacts that can encapsulate digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge"), but this phrase often carries a specific meaning with respect to semantic and linked-data approaches. Because of this, I've selected the phrase "computational research artifacts" and the abbreviated "research artifacts."

[5] "Packaging formats" refers to emerging conventions and standards for the dissemination of reproducible research artifacts such as [25, 53, 56, 140, 262].

As detailed in Chapter 3, these questions are addressed through multiple-case analysis [279]. I develop two primary and five supplementary cases, each representing a computational reproducibility initiative undertaken by a scholarly association, journal or conference. I define "computational reproducibility initiative" as a formal undertaking represented by policies, roles, workflows, artifacts, and tools intended to improve the computational reproducibility of published research. Case profiles are developed based on the semi-structured interviews of 17 key informants across the seven initiatives combined with qualitative coding and analysis [236] of documentary evidence and a sample of verified artifacts ($n = 27$).

1.2 Contributions and Key Findings

This study presents the first-of-its-kind analysis of how seven research communities are addressing concerns about research quality and rigor through the adoption of policies enforcing transparency and reproducibility of computationally derived results. In this section, I summarize the theoretical and practical contributions of this study as well as several key findings.

1.2.1 Theoretical and Practical Contributions

- **Framework for analyzing reproducibility initiatives.** I develop a novel framework for the analysis and comparison of the characteristics of reproducibility initiatives. This framework can be used to better understand the characteristics that affect operationalization of "computational reproducibility" through review and assessment processes.

- **Framework for analyzing reproducibility of research artifacts.** I develop a comprehensive set of dimensions of computational reproducibility that can be used to analyze and characterize both operational workflows and the resulting verified artifacts.

- **Expanding the research compendium concept.** I expand the concept of the research compendium as a mediating boundary object [246] between authors, editors, reviewers, tool developers, and repository operators.

- **Normative guidelines.** I develop a set of normative guidelines for computational reproducibility initiatives with implications for journal and conference leadership; tool and infrastructure developments; and funding bodies.

1.2.2 Key findings

In subsequent chapters, I present evidence in support of the following findings:

1. **Computational reproducibility is operationalized by initiatives through their organizational structures, policies, and review workflows (RQ1, Chapter 5).** While there are some similarities in initiative organization, there are significant differences in how policies are implemented that shape what they mean by reproducibility. All of the studied initiatives have implemented new editorial roles responsible for shepherding the assessment and all have also developed formal policies and assessment workflows. Key differences include whether the review is mandatory; who conducts the review; and what is actually reviewed or reproduced. In all studied initiatives, reproducibility review occurs post-acceptance and has no material effect on the acceptance decision.

2. **The core factors affecting computational reproducibility are common across the initiatives, but information requirements vary by type of computational research (RQ2, Chapter 6).** In Chapter 6 I identify four core factors including complete documentation of the computational workflow; accessibility of precise versions of software and data; sufficient information about the computational environment; and long-term accessibility of research artifacts. Information requirements vary in the case of research that relies on restricted access resources (e.g., protected or proprietary software, data, or hardware) as well as research that relies on large scale computational resources. Even for initiatives that mandate full reproductions are part of the review process, restricted access resources and large-scale computational processes require methods of assessing reproducibility without reproduction.

3. **There are material gaps in existing and available infrastructure (RQ3, Chapter 7).** Existing tools and dissemination formats present solutions to challenges faced by the studied initiatives, but are not yet widely adopted. Data repositories play a central role in the reproducibility initiatives, but lack capabilities (e.g., preserving the computational environment). Similarly, initiatives are constrained by limitations in editorial infrastructure.

4. **Changes to peer review are being used to correct misalignment of incentive structures and computational reproducibility practices of authors [RQ1, Chapter 5].** Earlier studies predicted that the increased availability of reproducible computational research artifacts would require policy changes by journals [90, 187]. Very low participation rates in the opt-in policies observed in this study further support the claim that current incentive structures in scholarly communication may even actively discourage self-provisioning of transparent and reproducible computational artifacts. The studied initiatives are leveraging existing incentive structures through an expansion of peer review to improve computational reproducibility practices of authors. However, initiatives must reconcile the high-cost of reproduction with the low value (i.e., information gain) of the review process. This is being achieved through 1) the use of students in full reproduction processes, 2) minimizing the information required for expert review (e.g., appendices), or 3) opt-in or invited paper policies.

5. **Journal reproducibility reviews require signifant human sources [RQ1, Chapter 5].** It easy to imagine that the reproducibility of computational research can be achieved with the "push of a button." The studied initiatives demonstrate that the implementation of formal processes to assess and verify computational research remains manually intensive. All seven of the initiatives have required new roles in order to conduct or recruit individuals responsible for the assessment process.

6. **Factors in community readiness may predict initiative success [RQ1, Chapter 5].** The studied initiatives represent the latest developments in decades-long community efforts to address concerns about research transparency and rigor. The success of these initiatives relies on substantial "cultural inertia" and a ready "installed base" [247]. While the most recent policy changes may be motivated by the "reproducibility crisis" narrative, they are largely made possible by groundwork already laid within the communities including both social and technical infrastructure. Leadership along with well-defined policies and workflows are required to make long-term changes to established submission and review processes that can survive editorial turnover.

7. **Gaps in infrastructure may impede widespread adoption of these practices [RQ1, Chapter 5].** The studied initiatives are driving changes to existing editorial and publishing infrastructure including editorial management, digital library, and repository platforms. There are substantial gaps required to support the review, publication, and discovery of verified artifacts. Investment in changes to editorial and publishing infrastructure may result in the increased adoption of reproducibility as-

sessment policies and workflows and, if implemented with measurement in mind, help to further our understanding of any effects of reproducibility assessment on research quality.

8. **General factors affecting computational reproducibility are discipline independent, but scope and scale vary [RQ2, Chapter 6].** The core factors involved in computational reproducibility are common across the studied initiatives. However, the level of detail required for some characteristics differs widely. For example, computational reproducibility requires detailed information about the computational environment that may include software, hardware, and network versions and configurations or even details of system runtime state. Even so, there is sufficient commonality across disciplines to suggest the need for a general set of guidelines to inform future policies. For research requiring large-scale, private or confidential resources, full reproduction may not be possible. Initiatives in many disciplines will benefit from guidance on policies and workflows handling the assessment of private, large, complex, and time-intensive computational artifacts.

9. **Advancements in reproducibility infrastructure will potentially reduce effort required by authors and reviewers. [RQ3, Chapter 7]** Many recent advancements in reproducibility tools and infrastructure have not had much impact on the studied initiatives. There is a clear opportunity for tools and infrastructure for computational reproducibility (e.g., containerization and virtualization, record-and-replay, automated provenance capture) to be applied in the service of reducing author and reviewer burden or increasing the transparency and reproducibility of published artifacts. Studies in cooperation with communities with high readiness could help to determine the potential impact of new tool and infrastructure on authors, reviewers, and future researchers.

10. **It is not proven that these efforts improve research quality and rigor. [RQ1, Chapter 5]** These initiatives are experimenting with changes to policies and incentive structures, but are generally not designed as such with no consistent means of measuring effect. It is commonly held but rarely tested that improved reproducibility would increase an authors citations or a journal's impact factor[6]. The experience from the field of economics discussed in Chapter 2 suggests that improved reproducibility should result in an increase in replications (for example see [32]). Future work would entail investigating whether these initiatives are more effective than other approaches

[6] One exception is the study reported in [264].

(e.g., pre-registration, education) or whether strict verification polices result in higher-quality artifacts than, for example, simple author checklists.

1.3 Organization

This dissertation is organized into 9 chapters, summarized here.

1. Chapter 2 reviews work related to the areas of computational reproducibility and associated infrastructure. I review the relationship between current initiatives and the broader landscapes of scientific investigation and scholarly communication, including peer review. I review the history of computational reproducibility including the "reproducible research," "replication standard" and "repeatability and workability" movements. This is followed by a look at the reproducibility crisis and how communities responded to related concerns. I also review technical infrastructure that has been developed to support improved computational transparency and reproducibility.

2. Chapter 3 details the methods used in this study including research questions, research design, theoretical frame, data collection, analysis, and study limitations.

3. Chapter 4 presents the detailed case profiles for the two primary and five supplemental cases.

4. Chapter 5 presents the results of the analysis of RQ1. The concept of computational reproducibility is complex and has different meanings to different communities. This chapter explores how the seven cases operationalize computational reproducibility in publication audit workflows. This includes the degree of reproducibility sought; the use of re-execution to ensure transparency and reproducibility; the expected skills of authors, reviewers, and verifiers; and the complexity of computational workflows used by authors. This chapter reports on the different verification models and the factors that influence their adoption. I conclude with a set of normative recommendations for communities considering the adoption of publication verification audits.

5. Chapter 6 presents the results of the analysis of RQ2. The adoption of publication reproducibility audits is intended to ensure that a published research artifact is reproducible to some degree. This chapter reports the results of an analysis of what makes an artifact reproducible (or irreproducible) based on the guidelines and policies

implemented in each of the seven cases. I conclude with a summary of the core factors involved in the assessment of reproducibility of computational results.

6. Chapter 7 presents the results of the analysis of RQ3. Each of the studied initiatives expects authors to make available their computational research artifacts using tools and packaging formats. This chapter presents the results of an analysis of initiative guidelines and policies and author practices as they relate to packaging artifacts for verification. I also review the features and capabilities of state-of-the-art reproducibility tools and infrastructure and how they can be used to meet policy requirements. I conclude with a summary of the elements required for the packaging and distribution of transparent and reproducible computational research artifacts. I present a conceptual model expanding the “research compendium” concept.

7. Chapter 8 concludes with a discussion of implications of the study and opportunities for future research.

Chapter 2

Background and Related Work

The study presented in this dissertation focuses on the very narrow concepts of "computational transparency" and "computational reproducibility." The concept of computational reproducibility has its origins in the publication and distribution of scientific software and data [128, 129], beginning in the 1960s; the "reproducible research" [40, 58] and "replication standard" [145] movements of the 1980s and 1990s; and the "repeatability" [174] movement in computer science in the 2000s. While early efforts in computer science, mathematics, and statistics focused on the publication, review and distribution of trustworthy and reusable scientific algorithms and software [128, 129], later efforts have also focused on the publication and distribution of data and programs used to support claims made in published research [40, 58, 145, 174]. In this chapter I review historical antecedents to recent computational reproducibility initiatives as related to broader concerns of scientific reproducibility and replicability; research production and quality; as well as scientific knowledge production and trust as they relate to the application of computational methods in research.

While the origins can be traced to decades past, the computational reproducibility movement of today has been fueled by the growing perception of a "crisis" in research reproducibility and credibility across the sciences [20, 28, 89, 244]. 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 [134]; study pre-registration [59]; changes in practice related to statistical significance [272]; and increased research transparency [67]. 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. As a result, many communities have begun enforcing new practices and policies for sharing the code and data used to support published research, with some of the earliest efforts in economics [266], political science [135], mathematics [122], and computer science [174, 156]. The origins of these recent initiatives can be traced to many of the earlier efforts with each community.

The use of the terms "reproducibility" and "replicability" in the limited context of computation has been the focus of much discussion [21, 125, 219]. In response to proposals to adopt computational reproducibility practices, opponents have often argued that they conflate the seemingly trivial reproduction of calculations, figures, and tables with the higher calling of scientific reproducibility and replicability [65, 125, 141]. Proponents often couch these initiatives as a "minimal standard" in the interest of transparency required to ensure research quality and integrity [144, 212].

Another dimension of computational reproducibility is the relation to the peer review and refereeing process. Journals concerned with the publication of algorithms or software have had specialized peer review since the 1960s [129]. With the rise of research data sharing in the 2000s, many journals adopted policies and processes for the review and curation of datasets associated with published research. The "reproducible research," "replication standard," and "repeatability" movements have all resulted in the creation and adoption of new forms of peer review and curation for journals whose primary focus is often not software. This has required the adoption of new editorial roles and structures and relied on the development of new types of infrastructure for scientific communication.

Over the past three decades, these efforts have resulted in a remarkable amount of infrastructure designed to support the creation, publication, and distribution of "computationally reproducible research artifacts," often in concert with broader technological developments. This includes tools in support of interactive analysis [133, 148]; reproducible documents [238, 232, 163, 278, 23]; provenance capture [218, 96, 56]; automation and workflow management [41, 71, 96, 101, 237, 273]; publishing as well as repositories [1, 147], packaging formats and related standards for software distribution [26, 105, 176]. The rise of virtualization and emulation technology [31, 158, 113, 262, 56] has made it possible to capture elements of the computational environment to further ensure the reproducibility of computational research.

In this chapter, I review the related literature to provide background for the reported study. I begin with a brief review of theories of science and scholarship as they relate to the broader concepts of reproducibility, replicability, and knowledge production. This is followed by a brief look at scholarly publishing and peer review practices, including recent practices related to sharing data and code. I review antecedents to the reproducible computational research movement followed by the events and responses to the "reproducibility crisis" and led to the reproducibility initiatives that are the focus of this dissertation.

2.1 Science, Trust, and Scholarly Communication

There is no one science, no single scientific culture, no method that dominates approaches to scientific knowledge production [44]. Broadly speaking, science is concerned with systematic observation and experimentation; the development and testing of theories and hypothesis; the sharing and confirmation of results within a community; and establishing trust in those results. How this is achieved varies widely across and even within disciplines and changes over time as research communities develop and enforce their own norms and standards for what constitutes "good" or "bad" practice. The variety of approaches to research are many: experimentation, simulation, or observation; quantitative or qualitative; theoretical or applied; grounded or hypothesis driven, just to name a few. However, even with this variation there is perhaps one constant: for the results of research to be accepted and built upon, they must be deemed trustworthy and published into the "scholarly record."

Trust is involved in all stages of scientific knowledge production. While scientific knowledge relies on evidence, that evidence is made available through a series of trusted relationships, of belief in ability and reliability [118]. As discussed by Darch [68], Wilholt [274] identifies two types of *epistemic trust* involved in science. The first is trust in the methods used to produce results and the second is trust in the researcher(s) who produced them. Trust can be found in the process of peer review, which relies on the integrity and ability of reviewers to assess scientific results and claims. Peer review remains one of the most important factors for determining the quality and trustworthiness of published research [261]. Researchers also trust journals, publishers, and repositories to maintain accurate records of scientific findings and implement appropriate publication processes including peer review with qualified reviewers.

Scholarly publishing defines the criteria that determine whether and how research outputs are accepted into the "scholarly record" and made available for use by others. It is the tail-end of scientific knowledge production with an immense cultural inertia, entwined a highly competitive publishing industry. Regardless of field of study, modern scientific and scholarly publication looks much the same. Authors conform to discipline-specific norms for writing, such as the IMRAD (introduction, methods, results, analysis, discussion) model, translating their research process and findings into a narrative structure. A draft of the paper is submitted to a journal or conference where editors consider the suitability of the paper based on various criteria and, if not immediately rejected, engage the mechanism of peer review. Through peer review, external reviewers provide feedback to the editors and authors and grade the paper according to some criteria with varying degrees of transparency and anonymity [260]. An accepted paper often undergoes revision prior to publication. The final paper goes through a production process, often including typesetting and proof-reading, before being published and accepted into "the record."

The role of scholarly publishing is central to academic career progression. Over the past two decades, advocates for rewarding other forms of scholarly output, such as data and software, have made inroads for consideration in hiring, tenure, and promotion. However, the publication of papers and associated measures of impact, such as citation, still remain central factors in career growth and the evaluation of an individual's (or journal's) research quality [260]. Since the 1960s, scholarly publication and peer review processes for some fields have expanded to include the review and publication algorithms, software, data, and more recently the complete results of computational research workflows [135, 266].

2.2 Trustworthy Software and Data

The fields of computer science and statistics began publishing and reviewing research software in the 1960s. The *Communications of the ACM* (CACM) established a new "Algorithms" editorial department in 1960 to make available "coded versions of algorithms to a wider audience for both pedagogical and reuse reasons" which later became the Collected Algorithms of the ACM (CACM) [128]. In 1975, the ACM journal *Transactions on Mathematical Software* (TOMS) was created to "expand the opportunities to publish important results concerning mathematical software and significant computer programs" [229]. From the first issue, *TOMS* established an "Algorithms Policy" and refereeing process along with the Algorithms Distribution Service (ADS)[245] for the publication of programs. At the time and still today *TOMS* accepted two broad types of submissions: 1) fundamental research papers on the analysis and critical evaluation of computer programs; and 2) practically oriented, concrete research and development in areas including linear algebra, polynomial manipulation, and non-linear programming. The "Algorithms Policy" [93] defined allowable languages (initially Fortran, Algol, and PL/1), criteria for contribution, and established requirements for documentation, copyright and testing and has been revised six times. Notably, the algorithm policy did not apply to research papers.

Beginning in 1967, the Royal Statistical Society journal *Applied Statistics* adopted a similar "Algorithms" section modeled after the *CACM* [128]. The *AS* editors proposed the creation of a library of tested algorithms for statistical work that could be built up, published, and maintained [192]. The *AS* published guidelines for authors [200, 201, 202] noting that "[a]lgorithms are published for two purposes; for direct use and for communicating computing method" and that "[a]dequate refereeing of an algorithm can entail much computer testing, so that refereeing may take longer than for a paper of comparable length." The process of algorithm review was different than conventional peer review and involved exercising the provide software.

The ADS provided a way for *TOMS* to distribute the complete source of programs without including them in publication. In the mid 1980s, with the advent of electronic mail, the netlib service [78] was developed as an alternative to the ADS as a means for distributing mathematical software, including the CALGO collection. A similar service, StatLib, was later developed for the distribution of statistical software [153]. Today, many distribution networks exist for different languages and platforms (e.g., Comprehensive R Archive Network (CRAN) [130]).

A related historical development in the distribution of scientific software, particularly in mathematics and statistics, was the emergence of a new class of interactive analysis and visualization environments. The software distribution networks described above provided access to high-quality scientific libraries to facilitate re-use, education, and the transfer of methods to practitioners. However, they were still largely out of reach for the average scientist [228]. A vision emerged of a set of computational environments that could be used to solve complex scientific problems "on human terms" – i.e., without requiring significant programming expertise [103]. Termed Problem Solving Environments (PSEs), these environments were envisioned to provide low-barrier, flexible, and extensible frameworks for scientific computing. Examples of PSEs in the mathematical sciences include MATLAB and Mathematica [230]. Similar examples from statistics include S/S-Plus [27, 45, 46], SPSS [193], Stata [115], and R [133]. Each of these environments provide researchers with access to high-level programming environments combined with a suite of high-quality libraries for interactive analysis and visualization. Many of them are extensible, enabling researchers to create and distribute custom packages that can be reused and applied by others.

Donoho and Stodden [79] view PSEs as a central component of reproducible computational research practice. Through open and extensible frameworks like R, researchers can develop and package new methods as libraries for distribution via CRAN, which can easily be used by others. Similarly, PSEs can be combined with "literate programming" practices [150] and publishing tools to bring together analysis and visualization with final manuscript preparation [232, 163], resulting in the type of "reproducible research" originally envisioned by Claerbout [58].

2.3 Reproducible Research, Repeatability and the Replication Standard

What we call today "computational reproducibility" has its origins in four different traditions. First, the early efforts in computer science, mathematics, and statistics toward the review and distribution of high-quality scientific libraries, discussed above. Second, the "replication standard" movement of the 1980s and 1990s in political science and economics, exemplified by the work of King [145]. Third, the "reproducible research" movement started in the 1990s by geoscientists Claerbout and Karrenbach [58] and more generally adopted in statistics [40, 211] and signal processing [165, 154]. Finally, the "repeatability" movement in computer science, started in the databases community [174]. Each of these antecedents provides an alternative view into what it means to share the data and code behind published research for the purposes of transparency, reproducibility, and replicability.

An early example of efforts toward computational reproducibility is the "replication standard" proposed by King [145]. Building on earlier work in economics, including the *JMCB* study [74] (see section on economics below), King proposed that authors should share the code and data behind published political science research in support of future replications. Authors should provide sufficient information for the evaluation and ultimately replication of their work. The "replication standard" underlies many initiatives in the political and social sciences, including the development of data repository infrastructure [147] and data and code sharing policies [2, 87, 135]. King originally advocated for the sharing of replication materials as part of the publication process, but stopped short of review or verification.

The phrase “reproducible research” was first introduced by Claerbout and Karrenbach [21, 58] 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.” They define reproducibility as “running the same software on the same input data and obtaining the same results” 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” [219]. Schwab et al. [238] later describe the resulting ReDoc system used in the Claerbout lab, which relied on the “make” command traditionally used to manage the compilation of software to “build” a paper based on the LaTeX authoring environment. They describe the key motivation as the inability of researchers to reproduce their own computations and the difficulty new students faced reproducing colleagues’ results. Claerbout and Karrenbach’s conceptualization is the basis of the reproducible research movement that includes Buckheit and Donoho [40], Gentleman and Temple Lang [105], informed the development of “literate” computing environments [163, 232], and informed new initiatives and policies in statistics[215, 211], econometrics [151], and computer science [61]. For the many researchers who rely on computation in their work, this idea was quite intuitive. By incorporating good software development practices into both the research and publication processes, it should be possible to completely reproduce all of the computations reported in the resulting paper.

The “repeatability” movement in the databases community similarly advocated for sharing data and code to support published claims while also encouraging “workability” in the interest of extension or reuse of published methods [174, 173]. As in other fields in computer science, even simple repeatability faces additional complexity, as the computational elements expand to include not only software, but compiler, hardware, and even network configurations. Computational reproducibility in computer science research is closely related to issues of numerical reliability [17] and reproducibility across hardware configurations and in parallel and distributed systems [120]. In a recent study exploring the reproducibility of studies benchmarking parallel fast Fourier transforms, [9] proposed a set of reproducibility classes: bitwise reproducible, numerical error bounds, statistically reproducible. They are faced with a fundamental problem in scientific computing in general – not all numerical results are repeatable. The issue of numerical reproducibility and high-precision computations in computer science can be traced back to the early work of Bailey [18, 17] in high-performance computing performance research.

Across disciplines, critics of computational reproducibility efforts focus on the increased burden placed on authors and reviewers and the limited value of reproducing computations to improving research quality. In political science, Herrnson's critique [125] of King's proposal rested primarily on three points. First, King's misuse of the term "replication" overstated what is actually a "verification" process which would do little to improve the quality of research. Second, that King's recommendations would negatively impact original data collection, putting too much burden on authors to share their data before they had the opportunity to fully exhaust research opportunities. Finally, the policy would place too much burden on journals and editors, lengthening and complicating the review process. Keiding's critique [141] of the more recent *Biostatistics* policy focuses on what he calls the "substantive context" of statistical analysis. He argues that research involving statistical analysis includes not only the computations but also the selection of a model that requires insight into the scientific problem. "[I]t ridicules our profession to believe that there is a serious check on reproducibility in seeing if somebody else's computer reaches the same conclusion using the same code on the same data set as the original statistician's computer did." Still, he sees the initiative as useful, if misdirected. In machine learning, Drummond [83] argues that, while the general practices are largely beneficial, reducing scientific reproducibility to such a narrow definition is potentially harmful and requiring the submission of all data and code counterproductive. He claims that these editorial policies undermine trust essential to the peer review process and will result in the accumulation of code of questionable value, increasing the load on reviewers. Instead, he argues, we should be increasing trust in reviewers and reducing their workload. He asserts that "careful reviewing by experts is a much better defense against scientific misconduct than any execution of code."

Still, the early efforts represented by Claerbout & Karrenbach and King in many ways foreshadow the coming crisis in scientific reproducibility. Claerbout & Karrenbach's innovation was intended to help their local lab improve efficiency and reproducibility of their work, and to ensure that publications were based on the actual results of their analysis. King was concerned about the quality and replicability of quantitative research the field. Nearly three decades later, their work underlies initiatives intending to transform the review and dissemination of computational research.

2.4 "Reproducibility Crisis" and Responses

In 2005, John P.A. Ioannidis published an influential and controversial paper entitled "Why Most Published Research Findings are False" [134]. Through a mathematical model and simulation, he demonstrated that for many study designs "it is more likely for a research claim to be false than true" and that "claimed research findings may often be simply accurate measures of the prevailing bias." In the following decade, several high-profile attempts to replicate important studies in drug research [222], cancer research [28] and psychology [59, 60] provided evidence that the results of previously published studies may not be as reliable or replicable as assumed. Around the same time, concerns about research quality and integrity grew further with highly visible replication attempts leading to the discovery of major errors [121, 203] and accusations of fraud and misconduct [13, 14, 15, 16, 39]. These events culminated in claims of a "reproducibility crisis" in science [20] and quickly spread to other fields, as communities attempted to assess the replicability of their own findings. Researchers began recommending changes to improve the quality and reliability of published findings including improvements to study design, larger studies, and a decreased reliance on p-values in reporting statistical significance [134]; preregistration and pre-analysis plans to reduce publication, reporting, and selection bias [60]; and stricter requirements for data and code sharing in support of transparency, reproducibility, and replicability [13, 14]. In this section, I review how the "reproducibility crisis" has affected the fields of economics, political science, computer science, and statistics.

2.4.1 Reproducibility and Replicability Studies

As discussed in the previous section, concerns of reproducibility and replicability in economics and political science research predate the "reproducibility crises" of the mid 2000s. Beginning in the 1980s and continuing today, researchers have conducted reproducibility or replicability studies, where they attempted to access the original materials and assess the reproducibility of published findings. Most studies have focused on computational reproducibility [74, 102, 179, 181, 182, 47] and journal policies [47, 74, 143]. A recent study by Berry et al. [32] attempts to assess true replication rates via citation analysis. Similar studies have been conducted parallel computing [42, 127], systems research [61], computational physics [254, 255]. The databases community has reported the results of conference evaluations [174, 173, 34].

2.4.2 Reproducibility in Computational Sciences

Heroux et al. [124] distinguish between "computational sciences" and "computing sciences." Computational sciences, they contend, "use computational modeling, simulation and data analysis as vehicles for scientific discovery" whereas "computing sciences" also "view computing itself as their primary research focus." Concerns about reproducibility in the computational sciences culminated in the organization of a series of workshops to explore factors and develop recommendations for scientists, journals, and funding agencies. These include the Yale Law School Roundtable on Data and Code Sharing and the resulting "Data and Code Sharing Declaration" [251]; the Vancouver Workshop at the Applied Mathematics Perspectives satellite conference to the International Congress on Industrial and Applied Mathematics (ICIAM) [166]; and workshop at the Institute for Computational and Experimental Research in Mathematics (ICERM) [252].

2.4.3 Reproducibility in Computer Science

Concerns about the quality and reproducibility of computational research permeated subfields of computer science including signal processing [22, 154, 264, 265], databases [174, 173, 34], systems research [269, 156, 157], and mathematical software [122]. Beginning in 2008, the ACM Special Interest Group on Management of Data (SIGMOD) began exploring the the adoption of review processes to assess the repeatability of research published at its annual conference [174]. Table 2.1 summarizes the participation and replication rates reported by the initiative. The first initiative was discontinued in 2012 due to low participation rates. The subsequent "db-reproducibility" initiative was in initiated in 2015 and is active today. In 2011 the systems research community began exploring the adoption of "artifact evaluation" (AE) for conferences. [156, 157]. Table 2.2 lists nine conferences that have adopted the AE process since 2011. While there are different approaches, the general AE process is becoming standard for reproducibility assessment in ACM and IEEE conferences.

In 2015, the ACM established a Task Force on Reproducibility, an initiative of the Publications Board, to work with ACM conferences and journals "to understand and articulate common best practices in preparing and reviewing artifacts, and how to reflect them in publication and enable their re-use." The task force produced the Artifact Review and Badging policy, first published in 2016. Also at this time, the ACM Transactions on Mathematical Software (TOMS) became one of the first journals to implement a comprehensive policy for computational reproducibility [122]. The journal added the optional Replication Computational Results (RCR) to the manuscript review process with the goal of providing

Year	Accepted	Participated	Repeated	Source
2008	78	54	29 (54%)	Manolescu et al. [174]
2009	64	19	10 (53%)	Manegold et al. [173]
2010	80	18	6 (33%)	`https://event.cwi.nl/SIGMOD-RWE/2010`
2011	88	34	24 (71%)	Bonnet et al. (2011) [34]
2015	94	10	10 (100%)	ACM DL, db-reproducible
2016	99	14	14 (100%)	ACM DL, db-reproducible
2017	94	8	8 (100%)	ACM DL, db-reproducible
2018	94	8	8 (100%)	ACM DL, db-reproducible

Table 2.1: Repeatability rates in SIGMOD conferences 2008-2018

Conference	Years
ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)	2011-2015†
European Conference on Object-Oriented Programming (ECOOP)	2013-2020†
ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)	2014 - 2020†
ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)	2014-15, 2018, 2020†
IEEE/ACM International Symposium on Code Generation and Optimization (CGO)	2015-2020†‡
ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)	2015-2019†‡
IEEE International Workshop on Visualizing Software for Understanding and Analysis (VISSOFT)	2015-17, 2019†
ACM SIGPLAN Object-oriented Programming, Systems, Languages, and Applications (OOPSLA)	2013-2020†
ACM SIGPLAN Symposium on Principles of Programming Languages (POPL)	2015 - 2020†

† http://evaluate.inf.usi.ch/artifacts
‡ https://ctuning.org/ae/prior_ae.html

Table 2.2: History of artifact evaluations in ACM/IEEE conferences from 2011 - 2020

"independent confirmation that the results contained in a manuscript are replicable."

2.4.4 Reproducibility in Parallel and High-Performance Computing

Hunold and Träff [132] revisit issues of trustworthiness and reproducibility in parallel and high-performance computing research. The authors argue that parallel computing research often leverages resources with many processors or cores and complex interconnections through shared-memory networks; compiler versions and settings matter. In high-performance computing, many of the systems have restricted access and a limited lifespan. Reproducibility may not be possible, as the original experiment was one-time only. In large scale parallel systems there may be additional interference from networks, file systems, or other users' workloads, which are very difficult to control. They consider both scientific and numerical reproducibility as prerequisites. The authors report on their experience using the VisTrails package and call for "an unbiased, scientifically sound survey whether experimental results in parallel computing are reproducible to our standards or not." The authors report the results of a preliminary survey on [131].

Hoefler and Belli [127] revisit the earlier critique of Bailey in their article "Scientific Benchmarking of Parallel Computing Systems: Twelve ways to tell the masses when reporting performance results." They analyze 120 articles from leading conferences including HPDC, SC, and PPoPP between 2011-2014 and present a twelve "rules" for improving community practice. These include guidance for summarization of results, statistical comparisons, experimental design, and results reporting. They argue that "[t]he complexity and uniqueness of many supercomputers makes reproducibility a hard task. For example, it is practically impossible to recreate most hero-runs that utilize the world's largest machines." They introduce the concept of interpretability as a weaker notion of reproducibility. "We call an experiment interpretable if it provides enough information to allow scientists to understand the experiment, draw own conclusions, assess their uncertainty, and possibly generalize results." While Hunold suggest parallel computing adopt Drummond's "scientific replicability," they simply advocate for "Clear documentation to ensure interpretability."

2.4.5 Reproducibility in Economics

In 1982, the National Science Foundation (NSF) funded a small study to look at the effect of journal policy changes on the replicability[1] of research published in the *Journal of Money,*

[1] The terms "reproducibility" and "computational reproducibility" are not common in the economics literature [21, 266]. The term "replicability" has generally been used to refer to a spectrum of activities. Pesaran [216] refers to replication in a *narrow sense* ("checking the validity of calculations or by carrying the

Credit and Banking (JMCB) [74]. The results of the study gave rise to a decades-long discussion of computational reproducibility and replication in economics, prompting changes in journal policies and even the development of theoretical models to explain author and publisher incentives to reproduce or replicate prior research. In July 2019, nearly four decades later, the American Economic Association (AEA) adopted a broad data and code sharing policy that embodies many of the lessons learned from the *JMCB* study and subsequent research.

The story of reproducibility and replicability in economics touches many aspects of the current discussion. This includes the practices and incentives (or disincentives) for authors and editors; the norms of the field related to peer review and research quality; sharing code and data; failed reproductions of the results of published papers; numerical reliability of standard software tools; difficulty of undertaking reproductions involving complex models; and finally the role of students in these initiatives. I've included this chapter because I believe that the lessons of economics still have much to teach other fields and, as I will discuss in later chapters, the AEA initiative presents a unique opportunity to understand the effect of computational transparency and reproducibility policies on research quality.

Inspired by the *JMCB* study, Mirowski & Sklivas [187] propose a game-theoretic model to explain incentive structures for replications in empirical economics research. They consider the "new entrant" into empirical research who has the choice to replicate or to reproduce

estimation...using other computer packages") or a *wide sense* ("if the substantive empirical finding of the paper can be replicated using data from other periods, countries, regions, or other entities as appropriate."). Hamermesh [117] distinguishes between *pure replication* ("duplicate, repeat, as in a statistical experiment"), *statistical replication* ("different sample, but the identical model and underlying population") and *scientific replication* ("different sample, different population, and perhaps similar but not identical model"). In many of the studies discussed in this chapter, "replication" is used to mean the same thing as computational reproducibility – in King's terms "running the same analyses on the same data to get to the same result, what should probably be called 'duplication' or perhaps 'confirmation.'" [145] Vilhuber [266] later adopts the definitions more closely aligned with NASEM from [33] where *reproducibility* refers to "the ability [...] to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" and *replicability* as "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected." In the discussion of economics, I'm stuck with the decision to use the proposed NASEM language to inaccurately reflect the concepts within a discipline or to continue with the ambiguous usage. I have decided to continue with true but ambiguous language of the field.

and extend earlier work. Reproduction and extension are easier because they have “no need or desire to understand the original result in all its gory details” while “encouragements are ubiquitous (grants, etc.), and the journals are predisposed to accept the new finding whether they are thought to limit the significance of the original report or to expand them.” They argue that the reproduction and extension is a welcome contribution as it cites previous work and legitimizes previous editorial choices. Replication is harder and less rewarding, because the “new entrant” must not only engage with the details of the original work, but claiming or proving that it is wrong creates additional problems:

> A successful disconfirmation always implicitly calls into question the refereeing competence of the journal, and also makes inordinate demands upon journal editors to adjudicate the inevitable controversy which ensues between the originator and replicator about the nature of asymptotically satisfactory replication.[2]

I refer the reader to the original paper for the model details. Mirowski & Sklivas conclude that, according to the model, “[r]eplicability will not be an observed activity unless journals set their information requirement high enough to encourage replicators.” However, setting the information level too high “would be attempting to duplicate what the entire social structure of science already exists to do, namely inculcate the new entrant with all the requisite tacit background knowledge to be able to attempt the replication.” The journal editor would likely set the information requirement high enough to “make it a respected journal” without placing undo burden on authors, driving them to publish elsewhere. Based on their analysis, they conclude that if an academic field wanted to increase replications:

> Among other issues, it would involve changing the incentive structure of empirical research; it might involve subsidies to replicators... to offset the cost differentials between replication and reproduction; it also might just involve going around the entire structure of costs and benefits by requiring apprentice empiricists (perhaps at the graduate student level) to attempt replication of one or more articles in the same way they are now required to do theses.

[2] As will be shown later in the case of *Political Analysis*, verification of computational reproducibility does nothing to change this and may in fact complicate it further.

Sample	Attempted	Fully replicable	Partially replicable	Study
154	35%	4%	4%	DTA (1986)
186	37%	23%		MMH (2006)
67	66%	37%		Chang Li (2015)
203	75%	14%	36%	Galiana (2017)

Table 2.3: Replication rates reported in four economics studies

As will be presented in later chapters, several of the policies in place today have done just this, including that of the AEA. In economics and other fields, the information required of authors has progressively increased as communities have explored different incentive structures to encourage and enforce replications of published papers. The more comprehensive policies, such as that of the AEA and *AJPS*, involve both subsidies and students, as students and professional statisticians are paid to verify the computational reproducibility of provided artifacts. The interplay between incentive structures for authors, editors, and reviewers are key to the initiatives that are the focus of this study. They each increase the information required of authors, change the incentive structures, and try to find economical ways to replicate studies – in many cases leveraging the well-suited characteristics of students or non-traditional researchers. However, unlike the original *JMCB* study, these initiatives are undertaken largely based on the intuition of community leaders and not designed as experiments.

2.5 Disincentives of Computational Reproducibility

Feigenbam & Levy [90] also explore the incentive structures for producing replicable empirical economics research, proposing and testing a theoretical model. Using the "narrow" definition of replication as "locating original data and verifying results from the reported computations," (i.e., computational reproducibility) they conduct an experiment to explore "the factors that impact the probability that data are provided with an article or made available by the author." The theoretical model considers that both researchers and editors have a similar objective function: maximize the citations to one's work or to the journal. Today, we can see this as maximizing the author's h-index and maximizing the Journal Impact Factor (JIF). Feigenbam & Levy argue that researchers risk loss of citation on an unreplicable article, but any radical devaluation would only occur in the case of fraud, not carelessness. Therefore the author has two options: work carefully to ensure results can be positively replicated or increase the costs on replicators by not providing materials upon request. They conclude that there is a powerful disincentive for untenured faculty as long as non-replication is not factored into promotion. Citations, on the other hand, are related to wealth-enhancement via salaries and promotions.

JIF-maximizing editors will benefit from an increase in citations from a perception of higher quality. However, the authors predict the probability that someone will try to replicate a particular study increases with the quality of the journal. Higher-quality journals are therefore incentivized to ensure the replicability of published work. The better the journal of the original article, the more citations a replication will receive, so a citation-maximizing researcher would produce replications when they are of lower cost to complete. Feigenbam & Levy argue that negative replications are closer to original contributions and that positive replications are generally bundled with new original contributions because of current incentive structures in publication (i.e., novelty). They conclude that:

> providing data offers an external monitor of quality and is simultaneously the best evidence of confidence in one's work. Without this external monitor, there are only internal costs and benefits of carefulness to consider. As we have suggested previously, these internal costs and benefits do not necessarily compel researchers to work of pristine quality.

Feigenbam & Levy present and test a model to explain author and editor incentives in undertaking and publishing replications. Their work is related to the "paradox of reproducibility": that published science is expected to be reproducible, but reproductions are rarely undertaken and journals are unlikely to accept reproductions, except when extended to present novel findings [43]. This is further supported by the later work of Berry and others [32] who in a study of "broad" replications of 70 empirical papers published in *AER* found 52 replications of which 44 (85%) had a wider scope (extension or robustness text).

With respect to the study presented in this dissertation, we might expect that any initiative that proposes to work within the constraints of current publishing and replication incentive structures would benefit from an understanding of the dynamics involved. Following Mirowski & Sklivas, to ensure replicability, editors will need to increase the information requirements on authors and find alternatives to current incentives to conduct actual replications. Following Feigenbaum & Levy, replicators will generally publish replications if 1) they are negative and target papers published in high-quality journals or 2) are positive and included as part of an extension. Based on the findings of Berry and others, it seems reasonable to expect that many replications conducted today are just one step a new entrant might take in the production of novel findings. It also follows that the availability of detailed replication materials would reduce the burden on the replicator and also be more likely to protect the journal from negative replications.

In the second part of their paper, Feigenbaum & Levy report the results of a small replication study and as with the *JMCB* study encounter a variety of errors in the provide materials: incorrect computations, unspecified data versions, mistakes by the replicator or original author, and numerical reliability of the underlying software.

2.5.1 Numerical Reliability and Reproducibility

> *Familiar to all is the tale of the researcher who solves the same problem using two software packages and obtains two different results.*
>
> – McCullough (1998)

The problem of numerical reproducibility is central to HPC research, but impacts other fields as well. Numerical reproducibility refers to round-off errors and other numerical differences that are "greatly magnified as computational simulations are scaled up to run on highly parallel systems"[252]. Bailey, Borwein, and Stodden [19] provide examples from the ATLAS experiment on the Large Hadron Collider, atmospheric simulation, and computational physics where the lack of numerical reproducibility across benchmark runs by the same team made it difficult to determine the sources of error, requiring expert analysis (see also [120]). Taufer et al. [259] report similar problems in molecular dynamics applications using graphics processing units (GPUs). Bailey and Diethelm [75] both highlight that, while numerical non-reproducibility is understood by the scientific computing community, users of affected methods and libraries may not. Bailey notes that bit-for-bit reproducibility is possible and that some applications have legal requirements, but that it runs longer and negates performance optimizations. Bailey also notes that many students who will go into technical computing fields lack rigorous training in numerical analysis. In a report to the NASEM committee, Bush (2018) reports on the state of reproducibility in climate science, where bitwise reproducibility is used for debugging. This is further described by [221].

In the late 1990s, McCullough [177, 178] and Vinod [181] published a series of papers assessing the reliability of standard software used in statistics and econometrics. McCullough proposed a methodology for evaluation [177] which was applied in [178], identifying multiple problems in popular packages, including SAS, SPSS, and S-Plus. In their discussion of identifying errors in econometrics packages, they return to the *JMCB* study, recommending that journal editors "require that authors identify their software (including version number) and make their code and their data widely available via archives." In addition to providing access to programs and code, such a journal policy could be used to identify results that were based on faulty software.

In this chapter, I have summarized the background and related literature for the reported study. The concepts of computational transparency and computational reproducibility are complex and deeply tied to norms and processes of scientific knowledge production. The initiatives that are the focus of this study are part of a broad effort across the sciences to improve research quality and integrity, and computational research specifically. While terminology is important, what will distinguish these initiatives is how they operationalize the concepts of transparency and reproducibility through policies and workflows.

Chapter 3

Methods

In this study, I explore practices related to the verification of computational transparency and reproducibility, which I refer to as *reproducibility initiatives*, and the packaging of digital artifacts that are generated as a result of these processes, which I call *computational research artifacts*. The primary method employed in this study is multiple-case analysis [279]. I develop two primary cases (political science and high-performance computing) and five supplementary cases (economics, statistics, biostatistics, mathematics, and databases). The primary cases, *AJPS* and *SC*, were selected because they are mature and likely to produce highly contrasting results. *AJPS* conducts verification through full reproduction while *SC* determines reproducibility through the assessment of information provided in an appendix. The five supplementary cases (AEA, *Biostatistics*, *IS*, *JASA-ACS*, and *TOMS*) were selected because they were likely to produce complementary results, enabling me to understand how representative the primary cases are and improve the generalizability of findings. The primary unit of analysis for each case is the reproducibility initiative. Each reproducibility initiative is represented by the organizational structure and roles; historical antecedents; documented policies and guidelines; operational workflows; and resulting research artifacts produced by authors and assessed by reviewers. A central part of this study is a set of semi-structured interviews with key informants [159] from the seven initiatives supplemented with qualitative analysis of verified artifacts and associated documentary evidence.

Through the in-depth exploration of these reproducibility initiatives, the multiple-case study provides a framework to compare and identify the factors that affect the creation, packaging, verification, and dissemination of reproducible computational research artifacts.

The multiple-case study approach supports the development of detailed conceptual models of the reproducibility audit process and packaging requirements for research artifacts as well as the development of normative guidelines and recommendations for communities seeking to undertake similar initiatives. In this chapter I review the study design including theoretical framework, research questions, case selection, sources of evidence, and analytical methods.

3.1 Study Design

This study is designed as a multiple-case analysis [279]. A preliminary theory of reproducibility initiatives was developed based on a broad review of the literature, drawing on concepts of reproducibility and replicability from the philosophy of science [118, 205, 225, 226, 277], social studies of science [62, 149], knowledge infrastructures [85], and information organization [246, 247]. After a survey of active reproducibility initiatives, a set of research questions were developed, case study approach defined, and cases selected. Each case was developed through a combination of semi-structured interviews with key informants and qualitative analysis of documentary evidence and published artifacts. Figure 3.1 presents a schematic of the overall research process. To develop the case reports, qualitative content analysis [236] was applied to the interview transcripts, documentary artifacts, and verified research artifacts. The resulting case reports are compared via cross-case analysis and used to develop a set of conceptual models of both the audit process and associated research artifacts. Relating back to the NASEM recommendations, these are used to develop a technical specification and a set of normative guidelines and exemplars intended to inform the development of publication audit processes, research artifact packaging formats, and related infrastructure.

3.1.1 Theoretical framework

The study design and analysis are informed by work from the philosophy of science [118, 205, 225, 226, 277], social studies of science [62, 149], knowledge infrastructures [85], and

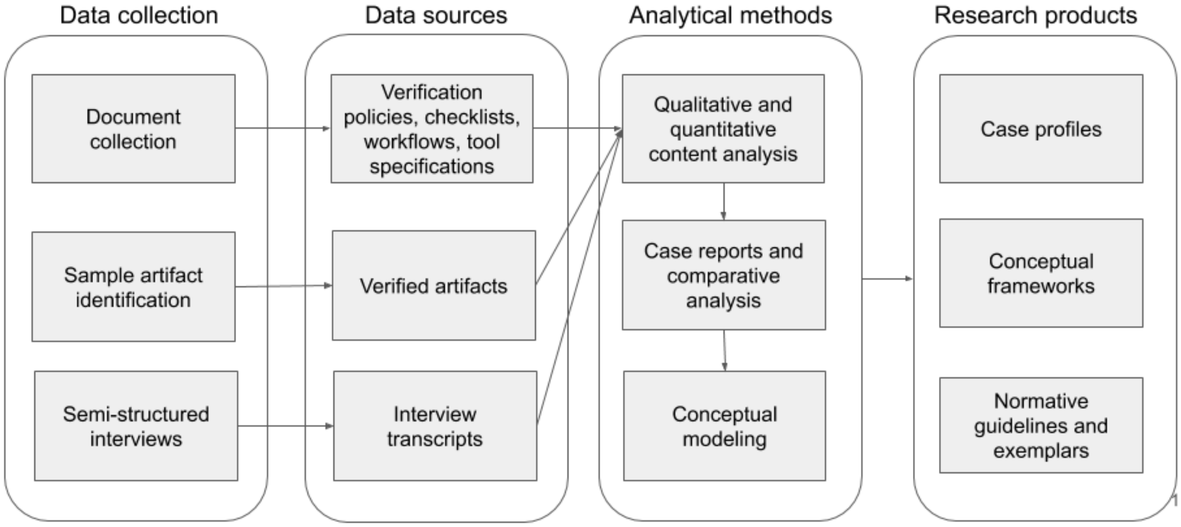

Data collection
Data sources
Analytical methods
Research products
Document collection
Sample artifact identification
Semi-structured interviews
Verification policies, checklists, workflows, tool specifications
Verified artifacts
Interview transcripts
Qualitative and quantitative content analysis
Case reports and comparative analysis
Conceptual modeling
Case profiles
Conceptual frameworks
Normative guidelines and exemplars

Figure 3.1: Study design overview

information organization [246, 247].

I take as a starting point Hans Radder's account of experimentation and reproducibility in the natural sciences [225] combined with the PRIMAD model of reproducibility in computer science [95] to better understand reproducibility of computational research as operationalized in the individual cases. The tools and infrastructure developed to support computational reproducibility through publication audit workflows and archiving of the resulting packages are understood through the lens of Edwards' "knowledge infrastructures" [85] and Star & Ruhleder's "installed base" [247]. The resulting packaged research artifacts that, through these initiatives, are an extension of the publication and the object of peer-review and archiving, are considered as mediating information objects, informed by Star & Grisemer's "boundary object" [35, 162, 246]. During the development of the supplementary case for economics, I also became aware of Mirowski & Sklivas's [187] and Feigenbaum & Levy's [90] theoretical frameworks for incentives in scholarly publishing, which further informs the cross-case analysis.

These theories provide the foundation for the development of the interview instrument, coding, case development, and cross-case comparison and analysis. Radder's typology of reproducibility and the PRIMAD model are the basis of questions relating to *what* is being reproduced by *whom* and what information is gained through the reproduction process. Knowledge infrastructures and the concept of the installed base informs my focus on historical antecedents and organizational characteristics of each initiative. The boundary object concept informs my focus on how different stakeholder groups view the role packaged reproducible research artifacts including researchers, editors, reviewers, re-users, and digital archivists. Economic theories of scholarly publishing were used during the analysis phase and informed my focus on the relationship between reproducibility initiatives and incentives of authors, editors, reproducers in the publication and peer review process. I detail each of these in the following subsections.

Radder's Typology

Hans Radder, a philosopher of science, provides an account of experimentation and reproducibility in the natural sciences that informs my understanding of computational reproducibility [225]. He introduces the concept of the *material realization* of experiments as the "experimental action and production either by the experimenters themselves or by laypersons" as distinct from the theoretical interpretation. According to Radder, while theoretical knowledge is required to prepare an experiment, it is not required to mechanically repeat the experiment. With this, he distinguishes three types of reproducibility:

1. Reproducibility of the material realization of an experiment under different interpretations
2. Reproducibility of an experiment under a fixed theoretical interpretation
3. Reproducibility of the result of the experiment by means of a set of different experimental processes

	Of What?		
By Whom?	*Reproducibility of the material realization*	*Reproducibility of the theoretical interpretation*	*Reproducibility of the result of the experiment*
By any scientist or any human being, in past, present, or future	1	5	9
By contemporary scientists	**2**	**6**	10
By the original experimenter	**3**	**7**	11
By the lay performers of the experiment	**4**	8	12

Table 3.1: Radder's reproducibility types and ranges

These types of reproducibility are further distinguished by different of levels expertise and the specific activities performed. In Table 1 from his book *In and About the World* reproduced here as Table 3.1, Radder considers the three types of reproducibility as they relate to the theoretical, material, and social expertise required to reproduce an experiment.

Any reproduction may be undertaken by individuals of varying expertise and at different times (*By Whom?*). The reproduction may rely only on the material realization (i.e., following the recipe independent of the theoretical interpretation); consider the material realization in the context of the theoretical interpretation; or attempt to reproduce the result of the experiment independent of the material realization or theoretical interpretation (*Of What?*). Radder also makes the distinction between the terms *reproduction* and *reproducibility* where *reproduction* refers to actual events in the past or present of reproducing an experiment and *reproducibility* is the (fallible) possibility of reproducing the experiment.

Referring to Table 3.1, Radder concludes that boxes 8 and 12 are empty by definition, since laypersons have no theoretical interpretation. Boxes 5 and 9 are most probably empty, as they would require stability of theoretical, material and social conditions over time. Box

1 is probably empty except for a few “mundane examples” (e.g., producing static electricity while combing hair). The remaining boxes 2, 3, 4, 6, 7, 10, and 11, he claims, can be related to historical experimental examples.

With respect to the current study, Radder’s typology provides a conceptual framework for exploring *what* is being reproduced and *by whom* in each case. Radder’s types 1 and 2 are closest to the NASEM definition of computational reproducibility. Radder’s type 3 is closer to the NASEM definition of replicability. Since I am focusing on computational reproducibility, I am most interested in examples that fit Radder’s cells 2, 3, 4, 6, and 7. Specifically, in this study I consider the role of theoretical interpretation (versus the material realization) and the expertise of the reproducer in publication reproducibility audits and the creation of research artifacts. While Radder’s typology is a useful abstraction for the exploration of how computational reproducibility relates to scientific reproducibility, it does not specifically address many of the dimensions of computational experiments. For this, I rely on the PRIMAD model discussed next.

PRIMAD

The PRIMAD model was developed during a 2016 workshop on the reproducibility of computational experiments as a framework to better understand “information gained from different types of reproducibility” in the field of computational research [95]. The authors identify six dimensions in the reproducibility of computational experiments: platform (P), research objective (R), implementation (I), method (M), actor (A), and data (D). Changing one or more of these variables, they claim, should result in new knowledge (i.e., information gain). Changing only the actor (person running the experiment) provides independent verification. Changing the execution platform (software and hardware stack) tests the portability of the experiment.

In Radder’s terms, the PRIMAD model accounts for *who* (actor), the theoretical interpretation (research objective, method), and characteristics of the material realization (platform, implementation, data). Following Radder's example of early experiment of boiling of liquid, one can envision a similar taxonomy.

A key problem when conceptualizing reproducibility in computational research is whether the computational environment is the instrument used to study other phenomena or whether the environment itself is the object of study. In many fields of research today, computational methods and implementations of mathematical and statistical models are applied to study external phenomena. In computer science, however, research may study any aspect of computation, from hardware to algorithms and beyond.

In a exercise testing the applicability of the PRIMAD model to LIGO gravitational wave workflows Chapp and others [49] note limitations of the model. Specifically, the unclear distinction between implementation and methods (when are changes in an algorithm sufficient to call it a new method?), effects of different actors within large research groups, and the roles of different datasets (e.g., input, intermediate, results). They suggest that "each field of science develop its own domain-appropriate refinement to PRIMAD."

In this study, the PRIMAD model is used to characterize the variables of experiments that are changed, if any, as part of community-specific reproducibility initiatives.

Knowledge Infrastructure, Boundary Objects and the Installed Base

The development of tools to support computational reproducibility; the implementation of publication reproducibility audits; and the dissemination, curation, and archiving of reproducible research artifacts require both technical and social infrastructure, not to mention the general apparatus of research in a particular field. Edwards introduces the concept of *knowledge infrastructures* as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" [85]. This unifying conceptualization informs my analysis on the relationship of reproducible research artifacts and emerging reproducibility-related infrastructure and activities.

Starr and Ruhleder [247] report on an ethnography studying a large collaborative infrastructure project for biologists. They note the dual and paradoxical nature of technology on organizational transformation as “both engine and barrier for change.” On the topic if infrastructure, quoting Monteiro, they argue that

> [i]nfrastructure does not grow *de novo*: it wrestles with the ‘inertia of the installed base’ and inherits strengths and limitations from that base. Optical fibers run along old railroad lines; new systems are designed for backward compatibility; and failing to account for these constraints may be fatal or distorting to new development processes.”

The reproducibility initiatives that are the focus of this study are imposing changes in both the social and technical infrastructure of computational research. The initiatives require involvement from numerous stakeholders in research and scholarly communications and are driving changes in the publishing and peer review process. They touch infrastructure and processes from publishing and editorial management; repositories and digital libraries; down to the tools researcher use to conduct their work. At the core of these initiatives are various concepts of how researchers should package and distribute the artifacts used to support the conclusions presented in published papers. Whether they’re called “artifact appendices,” “replication data sets,” “research compendia” or otherwise, they all serve the same function: to aid in the assessment and verification of transparency and reproducibility in the service of future replicability.

The packaged research artifacts that are the subject of publication reproducibility audits represent an extension to the paper. A variety of approaches have been proposed and adopted for the creation, management, and dissemination of these packages [105, 196, 53]. As an emerging standard, these packages can be seen as a type of boundary object – an artifact in the research process that is interpreted differently by individual actors but maintains a common identity [246, 35, 162]. They are created by researchers, assessed by reviewers, verified and archived by curators, and discovered by other researchers. These actors have not necessarily reached consensus on the purpose or form of the packages, yet are still able to implement complete review and verification workflows. The concept of the boundary object allows me to treat the package as an abstraction as well as a technical artifact of the research process that supports these different activities, meriting the recognition of it as a first class research object.

Economic models of incentives

As detailed in Chapter 2, the economics community has developed several models of incentive structures related to research replication and scholarly publication. Mirowski & Sklivas's [187] argue that increased replication rates will require journals to increase the information required of authors. Feigenbaum & Levy [90] argue that under current incentive structures, replications will occur primarily if they are 1) negative and target high-quality journals or 2) positive and included as part of a reproduction-as-extension. These models informed cross-case comparison and analysis, enabling me to consider how the studied reproducibility initiatives relate to incentives for authors and journals.

3.1.2 Research Questions

This study addresses three research questions concerned with how computational transparency and computational reproducibility are operationalized through publication reproducibility policies and audits and implemented in related tools, infrastructure, and information standards. For authors and journals to ensure computational reproducibility in published research, as stated in the NASEM recommendations, they must first understand what it means for published research artifacts to be reproducible and how to implement processes to assess and confirm reproducibility. Similarly, for the development tools, infrastructure, and related information standards, developers must understand the needs and requirements of authors and journals. Technological advancements may also enable new capabilities. The research questions are posed as follows:

- RQ1. How are computational transparency and computational reproducibility operationalized through publication reproducibility audits?

- RQ2. What are the characteristics of research artifacts that make them computationally reproducible (or irreproducible)?
- RQ3. What are the characteristics of tools and packaging formats that enable computational transparency and reproducibility?

RQ1 and RQ2 are addressed through interviews of key informants involved in the implementation of verification policies in the target communities including editors, conference organizers, reviewers, and archivists. The interviews were transcribed and coded for qualitative analysis. The results of this analysis are combined with qualitative content analysis applied to the documented policies and workflows adopted by each initiative and a sample of verified artifacts. RQ3 is addressed through the comparative analysis of the policies and guidelines of the initiatives with the capabilities of available tools and packaging formats and the analysis of author practices in a sample of five verified artifacts from each initiative. This will result in a taxonomy of characteristics of packaging formats currently used to represent computationally transparent or reproducible artifacts. This taxonomy, along with the results of RQ1 and RQ2, are used in the development of an abstract model, normative guidelines and exemplars. All qualitative analysis was conducted using ATLAS.ti (v8.4.4).

3.1.3 Case selection

As has already been discussed, across the sciences, research communities are exploring ways to improve computational transparency and reproducibility in published research. Reproducibility initiatives have been underway for years in political science [135, 2, 87], economics [188, 267], statistics [211, 98], mathematics [122], signal processing [265, 264], databases [173, 174, 55], machine learning [99], as well as high-performance and parallel computing [258, 204]. For the purpose of this study, I have selected the *AJPS* "Verification Policy" in political science and the Supercomputing (SC) "reproducibility initiative" in computer science and engineering as the primary cases. These initiatives were selected because of the policy and process maturity, the unique characteristics of the domains, and the availability of key informants, documentation, and related artifacts. For cross-case comparison, I have selected five additional reproducibility initiatives including those of the journal *Biostatistics*, the *Journal of the American Statistical Association – Applications and Case Studies* (JASA-ACS), ACM *Transactions on Mathematical Software* (TOMS), Elsevier's *Information Systems*, and the American Economic Association (AEA). These initiatives were selected because of the policy and process maturity and relationship to the primary cases. Focusing on these cases allows me to explore the similarities and differences between disciplines (social sciences, statistics, and computer science) as well as journals and conferences. There are of course many other related initiatives that were not included in this study due to limited access to stakeholders and the limited scale of this dissertation. These include the many conference evaluation initiatives within the ACM and IEEE (such as those represented by `https://db-reproducibility.org`, *https://ctuning.org/*, and *https://www.artifact-eval.org/*) as well as related initiatives in journals including *Transactions on Parallel and Distributed Systems* (TDPS) [204], *Political Analysis* [2], the *Quarterly Journal of Political Science* (QJPS) [87].

Case profiles are reported in detail in Chapter 4, but I will touch briefly on the primary cases here. The *AJPS* has had a mandatory pre-publication audit process in place since 2015 [135, 57]. The audit process is part of a decades-long effort in both political science and the journal itself that can be traced back to discussions in the American Political Science Association (APSA) beginning in the mid-1990s [145, 125, 249]. The policy has evolved many times over the years, reflecting the concerns of the community, changes in editorial incentives, and technological advancements. The *AJPS* initiative has informed and been informed by related work in economics, including the policies of the journal *AER* and AEA.

Supercomputing similarly began its reproducibility initiative in 2015. Over a five-year period, the incentives to participate have changed as policies have become increasingly strict.

Whereas the majority of quantitative political science papers rely on relatively simple computational processes, *SC* papers push the boundaries of high-performance computing, sometimes relying on boutique and leadership-class systems that are only available to a handful of researchers worldwide. This poses unique challenges to anyone who proposes to confirm the computational reproducibility of a particular study. Additionally, while journals often have months to review papers and associated materials, conferences like *SC* are under greater time constraints, providing reviews in a matter of weeks. *SC* is also part of the Association of Computing Machinery (ACM) and the initiative is informed by related efforts within the association.

3.2 Research Method

Multiple methods are available to address the above questions including surveys, ethnographies, and interviews. Surveys allow a wide reach in data collection, but require an understanding of existing practices in sufficient detail to develop a reliable survey instrument. Because the proposed study is one of the first of its kind and I am examining only the two primary and give supplemental cases, there isn't sufficient information available to develop a reliable survey instrument. Ethnography allows the researcher to engage deeply with a community for longer periods to develop a rich understanding of day-to-day practices. However, this method is most applicable for centrally located research sites and requires access and cooperation for an extended period of time. Given the distributed nature of the reproducibility initiatives and the number and heterogeneity of target communities, ethnography is infeasible. Qualitative interviews and analysis of documentary evidence and other artifacts strikes a balance between depth and leeway during the investigation process than surveys and can more reasonably be applied to a wider number of cases in a shorter time frame than ethnography. Interviews require the cooperation of and access to participants, which as noted above was one factor in the selection of cases.

In this study, I use qualitative interviews supplemented by document and artifact analysis to develop each case. The interviews provided additional context for the development and implementation of existing policies and workflows provided access to additional non-public documentary artifacts. Sampling is purposive, with a focus on stakeholders who can provide additional insight into key aspects of each reproducibility initiative.

3.3 Sources of Evidence

The primary data sources for this study are 1) qualitative interviews and transcripts; 2) documentary evidence; and 3) verified research artifacts.

Initiative	**Interviewees**	**Documents**	**Artifacts**
AJPS (Primary)	8	23	5
SC (Primary)	4	15	5
AEA (Secondary)	1	13	5
Biostatistics (Secondary)	1	7	0[†]
IS (Secondary)	1	6	3
JASA-ACS (Secondary)	1	7	5
TOMS (Secondary)	1	3	4
TOTAL	17	74	27

[†] Published artifacts for the Biostatistics initiative are no longer available.

Table 3.2: Summary of types of evidence for primary and secondary cases.

3.3.1 Interviews

I conducted semi-structured interviews of 17 key informants from the two primary and five supplemental cases. Three interviewees were involved in multiple initiatives. Interviewees included lead editors ($n = 3$); managing and associate editors ($n = 4$), conference chairs ($n = 4$), archivists ($n = 3$), and reviewers/verifiers ($n = 3$). The interview protocol is included in Appendix B.

The interviews captured informant perspectives about each initiative including motivations, workflows, benefits, expertise, metrics, challenges, and community responses. The semi-structure approach provided a set of questions to guide the interview with flexibility to explore additional topics.

Purposive sampling was used to select participants who could offer perspectives on the initiative. For the primary cases, this included both initiative leads and individuals involved in supporting roles. All participants were involved in the design, implementation, and/or operational aspects of each initiative and as such authors were excluded.

All interviews were conducted and recording using the Zoom videoconferencing service[1] and professionally transcribed using services provided by Rev[2]. Transcripts were imported into ATLAS.ti (version 8.4.4, MacOS) for qualitative coding and analysis, discussed in detail below. The high-level codes used for this analysis are included in Appendix C.

3.3.2 Verified Artifacts

Verified artifacts are the research artifacts provided by authors that have been assessed based on the documented policies, guideline and workflows. A sample of 25 verified artifacts was collected from the seven initiatives. Up to the five most recent artifacts were selected for each initiative, although not all initiatives had five artifacts. The complete list of artifacts used in this study for each initiative is provided in Appendix F.

3.3.3 Documentary Artifacts

Documentary artifacts include policies, guidelines, and workflows produced by each reproducibility initiative as well as editorials or related publications. The complete list of documentary evidence used in this study for each initiative is provided in Appendix D. The documented reproducibility audit policies and workflows provide an incomplete picture of a community's intentions and expectations with respect to computational transparency and reproducibility. In many cases, key stakeholders have written about the process in editorials

[1] https://www.zoom.com

[2] https://rev.com

or related publications [2, 211, 122, 87, 135, 258]. Even so, the documentary evidence often leaves unanswered questions about the motivations behind the initiative, lessons learned, and operational workflows. While author requirements and guidelines are generally publicly available, internal workflows and requirements for reviewers are not. As part of the interview process, participants were asked to provide access to non-public guidelines, workflow documentation, and related documentation for each initiative.

3.3.4 Analytical Approach

Qualitative analysis was used to address all three research questions. Transcripts and documents were imported into the ATLAS.ti qualitative analysis software. For each question, 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 was conducted via iterative coding using both inductive and deductive approaches. For each research question, an initial set of codes was developed based on the literature on reproducibility and knowledge infrastructures. Open coding was used to identify additional themes and ideas from the transcripts and documentary evidence. The final set of codes were selected for focused coding and are presented in Appendix C. Initial codes were tested and refined based on a pilot analysis on a subset of transcripts and documents.

Case profiles were developed from the results of the qualitative coding and include the following themes:

1. Initiative organization
2. Historical antecedents
3. Policies and guidelines
4. Technical infrastructure
5. Artifacts, identifiers, badges, and metadata

6. Initiative metrics

Using the coded evidence, cross-case analysis focused on the similarities and differences between the seven initiatives with respect to the codes and case dimensions. One goal of this analysis was to identify key or common factors that contribute to particular operational decisions across the initiatives as well as to identify possible explanations for the observed similarities or differences.

The results of the case analysis were validated using two methods: triangulation of data sources [208] and member checking. Combining interview transcripts with supplemental documentary and artifact evidence allowed me to consider consistency across actors well as in public (e.g., publication) versus private settings (e.g., interview). Case profiles were verified through member checking. Profiles were provided to interview participants to review for accuracy.

3.3.5 Human Subjects

The research approach for working with human subjects was determined to be exempt by the University of Illinois at Urbana-Champaign Institutional Review Board (UIUC IRB). See Appendix B for the exempt determination letter along with approved recruitment, informed consent, and interview protocol materials. Signed informed consent documents were collected for each participant.

3.3.6 Study Limitations

This study is conducted using a case study approach and qualitative data collection and analysis techniques. A common critique of qualitative research methods is the lack of generalizability. Findings and claims made about the cases in this study do not necessarily extend to other cases that were not included. I attempted to address this through the use of multiple case analysis, but generalizability remains a limitation. Other limitations include the small numbers of interviewees and sample artifacts. While interviewees were all key stakeholders, they do not necessarily represent the full scope of the initiative. The small number of artifacts analyzed is due in part to the imbalance of available artifacts across the seven initiatives. While several initiatives have tens or even hundreds of artifacts, three had fewer than five. As the number of artifacts increases across initiatives, the study could be expanded in the future.

3.3.7 Availability of Research Products

In addition to the information included in this dissertation, materials used for qualitative coding and analysis that are not protected under IRB will be made available via Dataverse `https://doi.org/10.7910/DVN/CKOGZM`. This includes the qualitative codebooks and associated ATLAS.ti projects.

Chapter 4

Case Profiles

4.1 Introduction

I develop a set of seven case profiles for a multiple-case analysis. Each case is a "reproducibility initiative," a formal activity undertaken by a journal or conference to improve the transparency and reproducibility of research published through their venues. Each initiative is represented by formal policies, guidelines and workflows as well as organizational roles and infrastructure established for the assessment of computational artifacts. This chapter presents the profiles of the two primary and five supplemental cases. All cases were developed based on data collected between October 2019 and March 2020 and capture the state of each initiative at that time.

The primary cases are the *American Journal of Political Science* (AJPS) and ACM/IEEE *Supercomputing* initiatives. As discussed in Chapter 3, these cases were selected because the initiatives are mature, with initial policies established in 2015, and because they were expected to be highly-contrasting. The *AJPS* is a top-tier journal in the field of political science while *SC* is a top-tier conference in the field of high-performance computing and the respective communities have different approaches to computational research. Research published by the *AJPS* tends toward long-tail or small scale local computing, while SC research tends to use computation scale. While both have adopted new peer review policies for the assessment of computational artifacts, the constraints and scope of each are quite different. The supplementary cases (American Economics Association (AEA), *Biostatistics*, *Journal of the American Statistical Association* - Applications and Case Studies (JASA-ACS), *Information Systems* (IS), and ACM *Transactions on Mathematical Software*

(TOMS)) were selected because they were expected to be complementary to the primary cases. Like the *AJPS*, the AEA initiative is focused on empirical social science research, however the initiative is association-wide instead of being limited to a single journal. The TOMS and *IS* journals face many of the challenges of *SC* with respect to computational scale and complexity, but without the constraints of conference timelines. Researchers in the *TOMS*, *JASA-ACS* and *Biostatistics* communities develop many of the methods and models that are employed by researchers in these other fields. Focusing on these cases allows me to explore the similarities and differences between disciplines (social sciences, statistics, and computer science) as well as journals and conferences.

This chapter presents summary case profiles for each initiative. Each profile is divided into 6 sections: initiative organization; history; policy and guidelines; technical infrastructure; artifacts, identifiers, badges, and metadata; and initiative metrics. Detailed operational workflows were documented for each case and included in Appendix E. Case profiles were developed based on the methods described in Chapter 3 including semi-structured interviews of key informants, analysis of verified research artifacts and documentary evidence.

4.2 The American Journal of Political Science (AJPS)

In 2015, the *AJPS* adopted a formal policy for the third-party reproduction and verification of computational results reported in all published quantitative research[1]. The new "Replication Policy" was an evolution of earlier policies adopted by the journal [184, 276] and part of a wider movement in the discipline [2, 87, 217]. The new initiative established a process whereby archive staff at UNC's Odum Institute for Research in Social Science[2] review

[1] The policy was expanded to include qualitative research in 2018

[2] A second aspect of the initiative established in 2018 includes verification of qualitative, non-computational research by the Qualitative Data Repository (QDR). While this is an important aspect of the*AJPS* initiative, it is out of scope for the current study.

and certify the accuracy of materials through professional curatorial review and by actual reproduction of computational results in the paper using author-provided artifacts by paid graduate students and professional statisticians [57, 135].

The roots of the policy – and concerns about replicability in political science in general – can be traced back to the 1994 APSA "Statement on Statistical Reporting, Archiving, and Replication" [249] and the resulting discussion initiated by King's proposal of the "replication standard" [125, 145]. The 2015 policy expanded on earlier requirements that authors share research materials [184, 276] adding a verification step to ensure that materials provided by authors could actually be used to reproduce reported results. The policy change was motivated by factors external and internal to the journal and the wider research community. External factors included the broader open science movement, funding agency requirements, and general concerns about scientific credibility and reproducibility. Internal factors include concerns about the quality and transparency of research exemplified by the 2010 Hatemi controversy [203] and the 2015 Data Access and Research Transparency (DA-RT) initiative [67].

While community response has been generally positive, the verification workflow has introduced new costs and challenges to both authors and editorial staff. The verification process requires additional human and computational resources and administrative infrastructure. With financial support from the Midwest Political Science Association (MPSA), the *AJPS* initiative has demonstrated an expanded role for research data archives in the curation and verification process. *AJPS* has also recently experienced unplanned leadership changes that demonstrate the challenges of maintaining continuity of these nascent policies.

The verification process was made possible through technical advancements enabling sharing of research materials, in this case the Dataverse platform [147], and collaboration with the archive at the Odum Institute. Odum archive staff are responsible for the implementation and operationalization of the verification process, including curatorial and computational steps. The verification of computational research requires staffing (graduate students and staff statisticians) and access to computational resources through Odum and the University of North Carolina (UNC).

4.2.1 Initiative Organization

The MPSA, a regional division of the APSA, is the funding body for the *AJPS* and responsible for contracting replication and verification services with the Odum Institute and the Qualitative Data Repository (QDR) at Syracuse. The MPSA holds one of the largest conferences in the field with over 5,000 presenters annually.

The APSA is a leading professional organization for political scientists, with over 11,000 members worldwide. The APSA publishes four leading peer-review journals and maintains guidelines for professional conduct including the APSA *Guide to Professional Ethics in Political Science* [8]. The APSA has 49 organized sections with 18 sponsoring publication of additional journals including *Political Analysis* (Political Methodology section) and *State Politics & Policy Quarterly* (State Politics & Policy section).

The *AJPS* is the flagship journal of the MPSA and consistently ranks in the top 5 political science journals by impact factor. The journal accepts research regardless of method, but the majority of research uses statistical or other quantitative methods [29]. The *AJPS* has a rotating editorial team currently serving four-year terms with lead editor(s), managing editor, associate editors and a 60-member editorial board. Beginning in 2019, the managing editor has the option to continue across editorial team changes. The journal is published by Wiley and uses the Editorial Manager platform for manuscript submission and peer-review. Since 2012, the *AJPS* has required authors to deposit replication materials associated with published articles into a journal-specific section of the Harvard Dataverse repository[3].

Staff at Odum are responsible for third-party verification and certification of replication materials submitted by authors to the journal, managing the operational aspects of curation and verification, including staffing. Additional methodological and technical expertise is

[3] `https://dataverse.harvard.edu/ajps`

available from the Odum institute. Computing resources used for the replication/verification workflow are provided by Odum IT and UNC Research Computing, including support services. Odum curators are full-time professional staff. Verifiers are advanced graduate students or staff statisticians with technical and methodological expertise. Harvard's Dataverse staff maintain the operations of the Dataverse system.

The *AJPS* initiative was made possible largely through the collaboration between then-editor William Jacoby and Tom Carsey, former director of the Odum Institute and prominent political scientist. As lead editor of *AJPS* from 2015-2018, Jacoby participated in both the DA-RT and the resulting Journal Editors' Transparency Statement (JETS) efforts. Carsey served as director of the Odum Institute (2011-2017), editor for the journal *State Politics and Policy Quarterly* (SPPQ) from 2010-2014, and was an active member of the DA-RT ad hoc committee and SPPQ signatory to the JETS statement. In 2013, he was awarded a Sloan/ICSPR challenge grant to explore the implementation of a data citation workflow in SPPQ. This pilot led to the creation of the *AJPS* verification workflow. Carsey died in 2017 and Jacoby stepped down abruptly amidst controversy in 2018 [92] leading to an unplanned leadership change for the initiative. Jacoby was replaced by interim editor Jan Leighley (2018-2019) and the current editorial team of Kathleen Dolan and Jennifer Lawless (2020-).

4.2.2 Initiative History

There are a number of historical antecedents and events related to the 2015 *AJPS* policy and to reproducibility and replicability in political science and the wider social sciences. The discussion about replication in political science can be traced back in part to the efforts in economics in the 1980s [3, 74] (see also Chapter 2). In 1994, the APSA Political Methodology section issued a "Statement on Statistical Reporting, Archiving, and Replication" [249] and *AJPS* implemented its first "Replication Policy" [184] requiring that all papers include a footnote indicating how readers could access the data and programs used in published research. In 1995, King proposed the concept of the "replication standard" which was discussed in a dedicated symposium issue of the journal *P.S. Political Science* [125, 145]. The replication standard conceptualization remains a lasting contribution and has informed the design of related infrastructure including King's own Virtual Data Center (VDC) [1, 146] and Dataverse [147] systems. In 2010, growing concerns about transparency and replicability led the APSA to appoint an ad hoc committee on Data Access and Research Transparency (DA-RT) to "provide guidance for instantiating these general principles in different research traditions" [168]. Jacoby claims that the *AJPS* "has gone farther than any other journal in implementing the DA-RT principles through our replication and verification policy" [135].

In 2012, the APSA revised its *Guide to Professional Ethics in Political Science* [8] including the statement that "researchers have an ethical obligation to facilitate the evaluation of their evidence-based knowledge claims through data access, production transparency, and analytic transparency so that their work can be tested or replicated." At the same time, *AJPS* revised its "Replication Policy" requiring researchers to deposit all materials in Harvard's Dataverse repository [276].

The TOP Guidelines were published in 2015 [194] and the new *AJPS* "Replication Policy" announced [135]. Far from an isolated initiative, *AJPS* joined other journals in the field that had already implemented strict verification requirements. The *Quarterly Journal of Political Science* (QJPS) adopted a replication policy in 2005 [87]. *Political Analysis*, the official journal of the APSA Political Methodology Section, adopted a replication policy in 2012 where materials deposited in Dataverse were reviewed by a graduate editorial assistant [2].

4.2.3 Policy and Guidelines

This section briefly summarizes the available policy and guideline materials at the time of the study. Copies of these documents are included as part of the study materials and listed in Appendix D.

The *AJPS* provides links to the Verification Policy[4] on the ajps.org website as part of the author submission guidelines. The policy page includes links to "Guidelines for Preparing Replication Files,"[5] "Quick Reference for Uploading Replication Files" as well as quantitative[6] and qualitative[7] verification checklists. Verifiers are provided with training materials and example reports via the Odum shared filesystem, which are currently private.

The verification policy states that the "corresponding author of a manuscript that is accepted for publication in the American Journal of Political Science must provide materials that are sufficient to enable interested researchers to verify all of the analytic results that are reported in the text and supporting materials." As detailed in the workflow described in Appendix E.1, data, programs, and documentation must be submitted to the *AJPS* Dataverse. The policy includes special provisions for qualitative research and restricted access data.

4.2.4 Technical Infrastructure

The *AJPS* maintains the ajps.org website that serves as the primary mechanism for communicating guidelines and policies to authors, including the Verification policy. The *AJPS* is published by Wiley and uses Elsevier's Editorial Manager to manage the manuscript submission and review process. Wilson's 2012 policy [276] established Harvard's Dataverse as the sole archive for journal replication materials. The Odum Archive, responsible for the implementation of the operational workflow, has developed a custom database to manage and track the curation and verification process ("Dashboard"). Because of gaps in infrastructure, the *AJPS* editorial team also uses a custom Excel spreadsheet to track manuscript verification status. Email is the primary mode of communication between the journal staff and Odum. The Odum shared filesystem is used to archive intermediate submissions, guidelines,

[4] https://ajps.org/wp-content/uploads/2019/03/ajps-replic-and-verif-policy-2-27-18.pdf

[5] https://ajps.org/wp-content/uploads/2018/05/ajps_replication-guidelines-2-1.pdf

[6] https://ajps.org/wp-content/uploads/2019/01/ajps-quant-data-checklist-ver-1-2.pdf

[7] https://ajps.org/wp-content/uploads/2019/01/ajps-qualdata-checklist-ver-1-0.pdf

checklists, and training materials. Computational resources including virtual machines, licensed software, and access to batch-compute resources are provided by Odum IT and UNC Research Computing.

4.2.5 Artifacts, Identifiers, Badges, and Metadata

The verification process results in both public (or access-controlled) and private artifacts. These include the final accepted paper published by Wiley, final verified materials published in the *AJPS* Dataverse, peer revew information in Editorial Manager, intermediate verification materials stored on the Odum shared filesystem, and verification reports and tracking information stored in the Odum Dashboard.

The *AJPS* assigns Center for Open Science (COS) badges[8] for open data and open materials embedded as JPEG images in the Dataverse record for each reviewed package. Odum and QDR add statements to the "Notes" section in the Dataverse record stating that the dataset underwent an independent verification process.

DOIs are assigned to the paper by Wiley, the*AJPS* publisher, and the dataset by Dataverse. A "Replication Materials" section is added to the publish paper both online and in print/PDF. This includes a link to the dataset DOI in Dataverse. COS badges are displayed in the online Dataverse record only. The Dataverse record contains the link to the paper in the "Related Publication" section. No links are provided to the verification policy or guidelines used for review from the article in Wiley or associated Dataverse record.

4.2.6 Initiative Metrics

The *AJPS* editor reports[9] have historically included a number of metrics related to the journal including the JCR 2- and 5-year impact factor, Google Scholar *h*-Index, submission rates and average editorial turnaround times, editorial decision rates, as well as top-cited

[8] `https://cos.io/our-services/open-science-badges`

[9] `https://ajps.org/editor-reports`

and frequently downloaded papers.

Jacoby, Lafferty-Hess, and Christian [135] also report the time added by the verification process and estimated hours per article. The Odum Dashboard tracks the duration of the curation and verification process and includes copies of all curation and verification reports (forthcoming). The Dataverse platform provides information about per-dataset downloads. The Wiley platform reports article citations publicly (downloads are available to editors only).

4.3 Supercomputing (SC)

In 2015, the steering committee of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis – also known as Supercomputing or *SC* – approved a new initiative to "increase the integrity of scholarly work being conducted in the high performance computing ecosystem" [7]. An "Advisory Group on Reproducibility to the *SC* Conference, ACM, and IEEE," chaired by Wilfred Pinfold, was formed early in 2015 to develop policies and guidance for the *SC* reproducibility efforts. The new "Reproducibility Initiative" was intended "to promote and support replication and reproducibility of computational results" in alignment with broader ACM initiatives [258]. As a first step, authors of papers accepted for 2015 conference were invited to submit a proposal for their work to be reproduced as part of the next year's Student Cluster Competition (SCC). The authors would be recognized for their participation and be awarded the new "Results Replicated" badge as part of the nascent ACM Artifact Review and Badging Policy[10].

This was just the first step in a multi-year initiative to promote reproducibility in the conference technical program. For the 2016 conference, authors could optionally include an "Artifact Description" (AD) appendix[11]. for consideration for the SCC Reproducibility

[10] `https://www.acm.org/publications/policies/artifact-review-badging`

[11] The appendix template used in SC17 is available from `https://sc17.supercomputing.org/submitters/technical-papers/reproducibility-initiatives-for-technical-papers/artifact-description-paper-title/index.html`

Challenge, a new competition challenging students to reproduce results of an *SC* paper. In 2017, the AD appendix became a requirement for consideration for the Best Paper and Best Student Paper awards. SC17 also introduced a "Computational Results Analysis" (CRA)[12] appendix for the assessment of research published using specialized or hard to access resources. Starting in 2019, the AD appendix became mandatory for all submissions and conference organizers introduced three new committees to the technical program related to the initiative. Prior to 2019, the *SC* appendix was derived from earlier work by the cTuning Foundation[13], working toward improved reproducibility in systems and machine learning research communities.

The *SC* reproducibility initiative is shaped by the many challenges of reproducibility and replicability in parallel and distributed computing [17, 19, 127, 131]; issues of numerical reliability and reproducibility [19, 120, 259] and by earlier work on reproducibility within the ACM and IEEE in areas including signal and image processing [154, 265], databases [94, 174, 173], systems research [61, 156, 269], and machine learning [99, 100]. Reproducibility in HPC research is also related to research in record-and-replay techniques [50].

The challenges faced by the HPC community with respect to reproducibility also impact the scientific fields that leverage HPC methods in their research. This includes challenges related to the use of parallel and distributed computing techniques; problems with access to specialized and short-lived hardware; dependencies on specialized scientific software (e.g., mathematical, MPI, etc); and non-determinism in execution due to system scale and complexity. While not unique to the HPC community, these challenges are particularly acute for the *SC* community.

As a conference, *SC* is faced with additional time and resource constraints. Unlike other fields, in computer science many of the premier research outlets are conferences, not journals. Conferences have much shorter timeframes for reviews (weeks, not months) and review

[12] For the 2018 conference, the AD appendix remained optional and the CRA was renamed to the "Artifact Evaluation" (AE) appendix.

[13] `https://cTuning.org`

committees can be organized just-in-time based on the number of submissions. Conferences like *SC* also rely on different editorial infrastructure than journals. Additionally, the SCC Reproducibility Challenge component of the initiative is unique in that organizers conduct in-person meetings to discuss candidate papers.

While the Supercomputing reproducibility initiative builds on the experience of other computer science subfields, its relationship with the SCC Reproducibility Challenge is quite unique. The SCC has presented an opportunity to explore effective mechanisms for reproduction and replication of research and requirements development, albeit through the selection of work exhibiting particular characteristics. The SCC is also seen as a testbed for new ideas prior to introducing them into the wider technical program. It is also an example of how students benefit from participating in the reproduction process.

4.3.1 Initiative Organization

Started in 1988, *SC* is one of the largest annual conferences of the HPC community. Each year, *SC* attracts over 5,000 participants in its technical program. For papers, the technical program has an average acceptance rate of 24%, with 64 accepted out of 288 submitted in 2018. Since 2007, *SC* has also included the Student Cluster Competition (SCC), an undergraduate educational initiative intended to further engage students in the HPC community [119].

Supercomputing is co-sponsored by the ACM Special Interest Group on High Performance Computing (SIGHPC) and the IEEE Computer Society. The manuscript submission and review process is managed using the Linkling service. Proceedings have been published in the ACM Digital Library. SCC student papers are published in a special issue of Elsevier's *Parallel Computing* [104, 275].

The ACM is the world's largest scientific and educational computing society. Founded in 1947, the non-profit professional association has over 67,000 members; 37 special interest groups; and publishes over 90 journals, transactions, magazines, and newsletters[14]. The ACM maintains a code of ethics and operates one of the field's premier digital libraries. The ACM is governed by over two dozen volunteer boards, councils, and committees, including the Publications Board that is responsible for the ACM Digital Library and publishing processes[15]. The Publications Board is "responsible for maintaining ACM's position as the preferred publisher in computing. The Board also envisions ACM as the "principal curator of publication data for the field." The Digital Library Committee is charged with developing overall strategic directions for the Digital Library publishing platform including new services and new features.

The Digital Library Committee organized the 3rd Workshop on Software, Data, and Reproducibility in Publication held on December 7-8, 2017 in New York City. The workshop generated a first draft of a Best Practices for Reproducibility in Computing Research. The DL Committee is responsible for the development of the Artifact Review and Badging policy[16].

The ACM Digital Library is one of the field's premier scholarly resources. The DL provides access to over 482,000 articles with 30,000 added in FY18[17]. In 2018, the DL migrated from an in-house publishing platform to Atypon's Literatum platform[18], which had been successfully adopted by the American Chemical Society, IEEE, SIAM, and MIT Press. One motivation for the move was to treat "data and code as first-class objects". The DL has a hybrid open access relationship with arvix.org.

SIGHPC is one of 37 special interest groups in the ACM. SIGHPC co-sponsors several conferences including SC, the Symposium on Principles and Practice of Parallel Programming (PPoPP), Platform for Advanced Scientific Computing (PASC), and the IEEE International Parallel & Distributed Processing Symposium (IPDPS)[19]. PPoPP and PA*SC* have also im-

[14] https://www.acm.org/binaries/content/assets/about/annual-reports-archive/acmarfy18.pdf

[15] https://www.acm.org/about-acm/boards-and-committees

[16] https://www.acm.org/publications/policies/artifact-review-badging

[17] https://www.acm.org/binaries/content/assets/about/annual-reports-archive/acmarfy18.pdf

[18] https://www.atypon.com/products/literatum/

[19] https://www.sighpc.org/for-our-community/hpc-events

plemented artifact evaluation processes.

SC is a large conference with a planning committee charged with communications, financial management, infrastructure, exhibits and local arrangements in addition to the technical program. For the purpose of this case, I am primarily concerned with the following aspects of the conference organizations.

1. Student Cluster Competition (introduced in 2007)
2. Reproducibility Challenge committee (introduced in 2016)
3. Reproducibility committee (introduced in 2019)
4. AD/AE Committee (introduced in 2019)

Since 2007, *SC* has included the SCC, an undergraduate educational initiative intended to further engage students in the HPC community [119]. Beginning in 2016, the SCC introduced the "Reproducibility Challenge" where students are tasked with reproducing a subset of results in a paper published during the prior year's conference. The SCC is viewed by organizers as a testbed for new ideas prior to introducing them into the wider technical program.

The *SC* Technical Program committee is responsible for conference programmatic areas including papers, tutorials, panels, workshops, posters, invited talks, and proceedings. Beginning in 2019, the Technical Program Committee added chairs for Reproducibility, the Reproducibility Challenge, and the AD/AE appendices. The Reproducibility Challenge (RC) liaison is the "main point of contact with the student cluster competition" and coordinates with SCC chair to make sure RC challenge application is in line with the technical program.

4.3.2 Initiative History

Historical antecedents to the *SC* reproducibility initiative include the early work of Bailey on performance studies [17]; related work within the ACM community, particularly related to artifact evaluation and review [99, 156, 173, 174] as well as the ACM *TOMS* RCR initiative [122]. Leaders in the *SC* community have been involved in key discussions and workshops including the Yale Roundtable [198], Vancouver Meeting [166], ICERM Workshop [252], and XSEDE workshop [136] on reproducibility in scientific computing. The *SC* reproducibility initiative is informed by other work in the ACM community, but has adapted to the specific needs and constraints of the HPC research community.

The *SC* AD/AE workflow (detailed in Appendix E.2) is based in part on earlier work in artifact evaluation for computer science conferences, pioneered by the 2011 ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE) [156] and continued today through in large part through the artifact-eval.org and cTuning.org initiatives. Artifact evaluation, or AE, generally refers to a post-acceptance evaluation process conducted by an independent review committee, almost exclusively in computer science conferences, to assess whether computational artifacts generate results that are consistent with the submitted paper [54]. Table 2.2 lists nine conferences that have adopted the AE process since 2011.

4.3.3 Policies and Guidelines

Beginning in SC19, as part of the submission process using the Linklings system, authors are prompted to enter AD/AE appendix information directly into the submission form. The form includes a link to the *SC* Reproducibility Initiative documentation on Github[20] including the Author Kit[21] used to collect system details. All instructions are integrated into the submission form, which is a major change from the earlier AD/AE LaTex template. The Linklings submission form also includes links to general ACM policies, the ACM Code of Ethics and Professional Conduct. The "Appendix Review Instructions" providing guidelines to reviewers is not available publicly but was obtained for this study.

[20] `https://github.com/SC-Tech-Program/SCreproducibility`

[21] `https://github.com/SC-Tech-Program/Author-Kit`

4.3.4 Workflow Summary

The complete workflow is detailed in Appendix E.2. The *SC* reproducibility initiative workflow has four components and operates over multiple years. In this workflow summary, SC-1 refers to the previous year's conference, *SC* to the current conference and SC+1 to the next year's conference.

For the current *SC* conference, the AD/AE appendices are reviewed for completeness and eligibility for the "Artifacts Available" badge by the AD/AE committee. Only those papers that have been accepted by the technical program are reviewed.

Prior to SC, the RC committee reviews those papers from SC-1 with complete AD appendices. Papers are reviewed for feasibility (including architecture, openness), applicability (something an undergraduate can complete in 2-3 months), and tie-in to the technical program, but not for quality or impact. Three papers are selected and the authors are interviewed at *SC* to determined willingness to participate. A single paper is selected for the SC+1 Reproducibility Challenge and announced during April or May preceeding the conference[22] In January following SC, the RC committee begins preparations for SC+1. The RC committee creates the challenge[23], confirms that the application runs as expected, develops grading rubric. At SC+1, SCC participants are given the challenge to reproduce a few figures from the a subset of the paper's results using a dataset that was not used in the original paper. Students write reports and in the past a subset were selected for publication in a special issue of *Parallel Computing*. The *SC* paper is awarded the "Results Replicated" badge and authors are recognized during the *SC* awards ceremony with a certificate of appreciation.

[22] For example, `https://sc20.supercomputing.org/2020/04/15/sc20-student-cluster-reproducibility-committee-chooses-benchmark-wisely/`.

[23] Note that this requires mapping state-of-the-art HPC research to something achievable by undergraduate students.

4.3.5 Technical Infrastructure

SC maintains the `supercomputing.org` website that serves as the primary mechanism for communicating guidelines and policies to authors, including information about the Reproducibility Initiative. A separate website is provided for each year (e.g., `scYY.supercomputing.org`). Information about the reproducibility initiative is provided mainly through guidelines, policies, and blogposts hosted on the website. The *SC* conference uses Linklings, one of several conference management tools used within the ACM. Conference papers are published in the ACM Digital Library, based on the Atypon Literatum platform (as of 2019). The *SC* organization began using Github to maintain website content for the technical program[24][25]. Due to limitations in the Linklings system to support the SC19 artifact review process, for SC19 custom tools were developed to track reviews and communicate with authors along with a database for determining badge eligibility.

The SCC initiative maintains separate infrastructure, including the program website [26]. SCC hardware is provided by commercial partners. Beginning in 2015, the Elsevier journal *Parallel Computing* published a special issue with SCC student papers.

4.3.6 Artifacts, Identifiers, Badges, and Metadata

The *SC* Reproducibility Initiative results in multiple artifacts. As of SC19, each paper published in the ACM digital library includes the mandatory AD appendix and optional AE appendix. For those papers assigned the "Artifacts Available" badge, related software and data artifacts must be published to an archival repository and linked via persistent identifier. Peer review information is stored in the Linklings system with supplemental artifact evaluation information, including badge eligibility, stored in a custom database maintained by conference organizers.

[24] `https://github.com/SC-Tech-Program`

[25] `https://github.com/Collegeville/sc-reproducibility`

[26] `https://www.studentclustercompetition.us`

Badges are assigned in the ACM Digital Library. For the AD/AE review process, only the "Artifacts Available" badge is assigned to eligible papers. Papers reproduced as part of the SCC Reproducibility Challenge are assigned the "Results Replicated" and "Artifacts Evaluated and Functional" badges. As of 1/28/2020, *SC* has awarded the following badges based on the ACM badging guidelines: Artifacts Available (86), Artifacts Evaluated and Functional (3), and Results Replicated (2).

For SC17, a single paper was assigned all three badges. For SC18, all papers that included the AD appendix were assigned the "Artifacts Available" badge[27]. For SC19, only those papers where "associated artifacts have been made permanently available for retrieval. Author-created artifacts relevant to this paper have been placed on a publicly accessible archival repository. A DOI or link to this repository along with a unique identifier for the object is provided" were assigned the badge. This means that the policy for badge assignment changed over time.

No information is provided about which papers have appendices, aside from the full-text of the paper itself.

4.3.7 Initiative Metrics

The *SC* conference does not report any official metrics aside from acceptance rates. The following metrics were reported for SC19 [28]:

- Number of submissions (338), rejections (96)
- Number of appendices reviewed (242) and average per reviewer (31)
- Appendices completed (139), incomplete (87), missing (16)
- Number of email conversations with authors (107)
- Number of candidates for Best Paper (8) and Best Student paper (6) lacking appendices

[27] `https://sc18.supercomputing.org/submit/sc-reproducibility-initiative/`

[28] `https://github.com/SC-Tech-Program/SCreproducibility/blob/master/AD-AE-Appendices.md`

Unlike journals, conferences do not have a widely accepted equivalent of the JCR impact factor.

4.4 American Economic Association (AEA)

In 2019, the AEA announced a new Data and Code Availability policy that applies to six journals published by the association including the *AER*, one of the most prestigious journals in the field of economics [267]. This followed the creation of a new Data Editor position in 2018, a member of association leadership intended to be responsible for the creation and implementation of the new policy in cooperation with AEA journal editors and Publications Office [84]. The policy change included clear requirements for both code and data, the creation of a new archival repository, and staffing for the Data Editor to review and possibly re-execute all provided code to certify compliance with the new policy. The Data Editor, based at Cornell University, has defined an operational workflow that relies on graduate and undergraduate students as well as institutional computational and licensed software resources[29]. The AEA journals use Clarivate's ScholarOne for the manuscript submission and peer-review process, which is insufficient for the new workflow. The Data Editor's workflow is implemented in part using an instance of the Atlassian tools suite including JIRA for issue management and Bitbucket for source control. The AEA's ScholarOne instances were adapted support the new editor position and a conditional accept state. The AEA has partnered with the OpenICSPR repository for publishing data and code packages previously hosted as supplemental information on the AEA websites. As part of an earlier initiative, the AEA established a registry for randomized controlled trials (RCTs)[30]

[29] `https://github.com/labordynamicsinstitute/replicability-training/blob/master/jira-workflow-training.md`

[30] While not directly related to the reproducibility review process, the RCT registry is an example of another effort within the AEA to improve research rigor and quality.

4.4.1 Initiative Organization

Founded in 1885, the AEA is a non-profit scholarly association "dedicated to the discussion and publication of economics research while promoting the understanding of economics."[31] The AEA has over 20,000 members from academia, business, government, and consulting organizations and publishes 8 journals along with the proceedings of the association's Annual Meeting. The AEA retains complete control over the publication process, including publishing its own journals and providing online access via aeaweb.org through the AEA Publications Office.

In addition to aeaweb.org, the AEA *Papers and Proceedings*[32] is published yearly containing selected papers presented at the annual meeting along with reports from association leadership, notes from the executive committee, and reports from individual journal editors. The AEA publishes the following journals: *American Economic Review*, *Journal of Economic Literature*, *Journal of Economic Perspectives*, *American Economic Review - Insights*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Microeconomics*. The AER has had a long-standing policy for depositing data and code associated with published papers, initially adopted in 1986 [10] and revised in 2005 [30] and 2008 [189]. The *AEJ* area specific journals adopted the same policy beginning in 2009. In 2016, prior to establishing the Data Editor position, the AEA adopted a Data Availability policy across all journals based on the AER policy.

In 2017, AEA announced the new Data Editor position that would be responsible for the association's data policy for all journals going forward. The position was filled in 2018 by Lars Vilhuber, Executive Director of the Labor Dynamics Institute (LDI) at Cornell University [84]. The LDI is part of the ILR School, a leading applied social sciences college with a focus on work, employment, and labor policy issues. LDI works with research networks

[31] `https://www.aeaweb.org/about-aea`

[32] `https://www.aeaweb.org/journals/pandp`

and statistical agencies to provide access to novel data on the dynamics of labor markets. AEA review activities are operationally implemented in this organization. CISER provides research and computing services to social scientists at Cornell. For the AEA initiative, CISER provides access to specialized compute resources, licensed software, and IT expertise.

The ICSPR at the University of Michigan operates the OpenICSPR repository platform used by the AEA. OpenICPSR enables researchers to freely self-publish data as part of the ICPSR catalog [169]. OpenICPSR includes support for domain-specific metadata, optional review by social science librarians, and the ability to disseminate sensitive and restricted use data.

4.4.2 Initiative History

As discussed in Chapter 2, the origins of computational reproducibility efforts in the social sciences can be traced to early initiatives in economics. Economics researchers have for decades considered the question of research reproducibility and replicability [3, 32, 48, 74, 102, 179, 180, 181]. Many of these studies have included AEA journals and the AEA has been responsive to community criticism, adapting policies to address concerns about research quality [10, 30, 189].

The *JMCB* study was designed for "the collection of data sets and programs used in selected empirical articles in the journal... to evaluate findings through replication of empirical results." The authors conducted an experiment to try to replicate studies published in the journal before and after the adoption of a new policy requiring researchers to make data and programs available. In the first part of the study, they requested data and programs from the authors of papers in three groups: 1) those published before the policy change; 2) those accepted but not in print; and 3) those currently under review. From these three groups, they received data and code from 34%, 72%, and 78% respectively. In the second part, they reviewed the 54 packages received from authors to determine eligibility for replication and attempted replications on the eight that qualified (15%). Of these eight, they were able to fully replicate results from only two, while two produced “qualitatively similar” results, and one had programming errors that, when corrected, did not change the conclusions of the paper. Of the remaining papers, two could not be replicated – even with the help of the authors – and one cited source data that could not be found. They pursued a ninth replication of a large-scale econometric model that required significant technical expertise and, while successful, recognized that “formidable difficulties exist for studies based on large-scale models.” In the end, they were able to fully reproduce the results from 4 out of the 54 papers where authors provided code (7.5%) or of the 154 papers accepted by the journal (2.3%). These findings supported their recommendations that journals should require the submission of programs and data at the time papers are submitted. The *JMCB* implemented the policy as part of the study and other journals revised policies after the results were published. The *American Economic Review* (AER) published its first policy in in 1986 as a direct response to the *JMCB* results [10].

Years later, Anderson and Dewald [3] lamented the lack of adoption of their recommendations by other journals and note that the *JMCB* itself discontinued the policy in 1993. Reporting on a similar initiative at the Federal Reserve Bank of St. Louis (FRB), they note that the download rates for data and programs from articles at the FRB in the first year nearly matched those of the 10-year period at *JMCB*. They attribute this to a key technological change: in 1993 the FRB data was made available by “electronic bulletin-board” while for the *JMCB* authors had to mail in requests for data. Technology, they argue, had reduced the cost to authors of requesting the data below the “marginal value to an individual researcher of replicating a previous study.” Changes in technology for the storage and distribution of research materials may change the equation for replications.

In a subsequent study, McCullough and Vinod [182] repeat the *JMCB* study and attempt to replicate all empirical articles in a single issue of *AER* using the authors’ original code and data, which at the time was under the 1986 policy. Out of ten papers, they were only able to obtain complete materials from one:

> Regrettably, we had to abandon the project because we found that the lesson of William G. Dewald et al. (1986) has not been well-learned: the results of much research cannot be replicated. Many authors do not even honor this journal's replication policy, let alone ensure that their work is replicable. Gary King (1995, p. 445) posed the relevant questions: "[I]f the empirical basis for an article or book cannot be reproduced, of what use to the discipline are its conclusions? What purpose does an article like this serve?"

They conclude that their results further support Feigenbaum & Levy's conclusion [90]. The disincentives for authors to participate in replication of their work are too great. Nearly twenty years later, their work to study the numerical reliability of software used in published research led them to the same conclusions as Dewald, Thursby, and Anderson: to support replication, journals need to archive the programs and data used by authors.

McCullough and Vinod's findings had a near immediate impact. In a March 2004 editorial, then *AER* editor Ben Bernanke announced that the journal would pursue "more active enforcement" [30]. At the time, the *AER* had just completed a transition to an online manuscript management system and was establishing an archive at Duke University [63]. In 2005, the journal announced a new "Data Availability Policy":

> It is the policy of the American Economic Review to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the Review, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the AER Web site...As soon as possible after acceptance, authors are expected to send their data, programs, and sufficient details to permit replication, in electronic form, to the AER office.

The *AER* moved quickly to address the concerns raised by McCullough and Vinod and the technology was in place do it. Future studies would repeat the pattern: as authors continued to identify problems replicating previously published studies, *AER* and eventually AEA enforcement tightened, with increased information requirements on authors, new editorial infrastructure, and eventually increased requirements on reviewers.

In 2006, McCullough, McGeary, and Harrison [179] return to the *JMCB* archive to assess replicability of papers from 1996-2002. Of 150 empirical articles, they found that only 69 contained sufficient information to attempt replication and of those only 14 were successful (22%). This is compared to the 4% found by the original *JMCB* study. As a result, the authors argue that the field still needs more replications and that journals should have stricter archiving policies. In 2008 [180], they expand their study to include four other journals in the field and conclude that:

> Making sure that one's results are replicable is an enormous amount of work and, since no one is checking, the rational economist will not invest the amount of time necessary to ensure that his results are replicable. The journals do not check because they do not want to run the risk of admitting that they published irreproducible results. Moreover, the vast majority of journal editors simply is not willing to do what it takes to ensure that they are publishing replicable research in the first place.

They make a series of recommendations for an effective journal archive, stopping short of verification. The recommendation reads like many recent policies. They also note a key weakness of data-only archives, such as those of the *Journal of Applied Econometrics* (JAE), again citing King:

> The primary purpose of an archive is not to ensure replicability (King 1995, 494) but to enhance extensibility (which presumes replicability). Thus, an archive should make it easier for one researcher to build on the work of another, and part of this 'building' is, of course, being able to reconstruct (replicate) what the first researcher did. In this regard, any data-only archive fails miserably.

In 2015, Chang an Li [47] reported the results of yet another replication study followed by Galiani, Gertler and Romero [102] in 2017. In each case, the authors demonstrate that the archiving policy still isn't sufficient. The materials provided by authors, it seems, need to be verified.

The work of Gary King, a political science methodologist, is often cited in these studies. The interplay between economics and political science on the issue of replication began with King's proposal of the replication standard [145] and continues today. King cited the *JMCB* study as "an excellent example of a recent study of adherence to the replication standard." As can be seen in the quotes above, in return his paper is cited frequently in support of strengthened journal policies in economics. Shortly after these policy changes at the *AER*, political science journals began experimenting with the verification of materials prior to publication. In 2015, the *AJPS* announced a strict verification policy, whereby every accepted paper must go through a curation and replication process prior to publication [135]. The *AJPS* policy and workflow informed later work at the AEA [267].

At the 2017 AEA Annual Meeting, a search committee was appointed to fill the role of an association-wide Data Editor [84] who's role was intended to:

> design and oversee the AEA journals' strategy for archiving and curating research data and promoting reproducible research. In this capacity, the Data Editor would serve as a liaison between the journal editors, authors, and data custodians.

The position was filled by Dr. Lars Vilhuber, executive director of the Labor Dynamics Institute (LDI) at Cornell University. Vilhuber established the new "Data and Code Availability Policy" and workflow for the pre-publication review and verification of data and code submitted by authors of papers across the AEA's journals. Vilhuber establishes the goal of the initiative as *transparency*: "[t]o ensure the credibility of the scientific endeavor, transparency of the methods and data used are critical" [266]. Leveraging the resources of the LDI and Cornell University, the AEA has implemented what may be one of the most ambitious policies today.

4.4.3 Policy and Guidelines

The AEA provides a link to the "Data and Code Availability Policy"[33] during the submission process through ScholarOne and on the `aeaweb.org` website. The policy document includes links to deposit instructions for OpenICPSR[34], frequently asked questions[35], and unofficial guidance from the AEA Data Editor[36]. Verifiers are provided with training materials[37], including example replication reports, via Github.

The policy requires "[a]uthors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs." Data, programs, and documentation must be submitted to the AEA Data and Code Repository unless access is restricted or limited. The policy includes a special provision for restricted-access materials and extended rules for experimental work and RCTs.

4.4.4 Technical Infrastructure

The AEA maintains the primary association website, aeaweb.org, which includes association content and provides online access to all AEA journals. The website also serves as the primary mechanism for communicating guidelines and policies, including the Data and Code Availability Policy. The AEA Publications Office is responsible for the publication of and access to all association journals.

The AEA uses Clarivate's ScholarOne, a commercial tool commonly used to manage the manuscript submission and peer review process. Since ScholarOne lacks many of the capabilities required for the data review and verification workflow, AEA licenses the Atlassian

[33] `https://www.aeaweb.org/journals/policies/data-code/`

[34] `https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea-guidance.html`

[35] `https://www.aeaweb.org/journals/policies/data-code/faq`

[36] `https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea-guidance.html`

[37] `https://github.com/labordynamicsinstitute/replicability-training`

tool suite using JIRA for issue management, tracking and assignment and Bitbucket for in-progress review of materials. A custom email-based API was developed to automate JIRA issue creation based on ScholarOne referee assignment.

All computational resources required by verifiers including access to cloud- and batch-compute resources as well as licensed software[38] are provided by Cornell University through the Cornell Institute for Social and Economic Research (CISER) and Cornell Institute of Biotechnology (BioHPC). The AEA Data and Code Repository uses the OpenICSPR platform[39] developed and hosted by ICPSR at the University of Michigan. The AEA RCT Registry is based on software developed by MIT.

Github is used to host documentation and training materials. Email lists are used by the Data Editor to communicate with the verification team. The Data Editor also has a Twitter handle used for social media activities.

4.4.5 Artifacts, Identifiers, Badges, and Metadata

The AEA workflow results in both public (or access-controlled) and private artifacts. These include the accepted manuscript published by the AEA, final verified materials published in OpenICPSR, peer review information in ScholarOne, intermediate verification materials in AEA Bitbucket, verification reports in AEA Bitbucket, and issues in AEA JIRA.

There are currently no badges or related metadata in use by AEA. DOIs are created for the online version of the paper and the OpenICPSR dataset. Links are created between the published dataset and the online record. The published journal article or PDF includes the DOI for the online article record.

[38] Some licensed software used by authors may not be available for the review process.

[39] Previously, supplemental materials were hosted on the `https://aeweb.org` system. In addition to being an archival repository, OpenICPSR includes support for sensitive information and enables broader dissemination of research artifacts [268].

4.4.6 Initiative Metrics

AEA journal editors have historically reported a number of metrics related to the publication process. These include manuscript submission rates; revision outcomes; decision times for manuscripts, average processing times; distributions of publications by type and subject matter; and data posting policy conformance [108, 109, 110]. Although not reported, additional conventional metrics include the JCR 2-year and 5-year impact factor. The OpenICPSR repository provides usage metrics including views, downloads, and citing publications (for the dataset). The most recent Data Editor report [268] includes information about software used by authors; the number of assessment per journals; distribution of the number of "rounds" or resubmissions per manuscript; and the amount of time required for the assessment process. The report also includes information about the pre-registration of randomized controlled trials in the RCT registry.

Additional metrics can be found in the literature related to replications in economics. These include survey results of requests to authors for data, circulation rates, and counts of replications of individual papers [117, 86]. Currently no metrics are reported based on the AEA Data and Code Availability policy.

4.5 Biostatistics

In 2009, the journal *Biostatistics* announced a new policy [211] intended to encourage authors to publish "reproducible research"[40] and introduced the new role Associate Editor for Reproducibility [76, 211]. *Biostatistics* was one of the first journals to implement this type of policy and editorial role. The initiative built on the experience of initial Peng and Zeger in epidemiology research [210, 214, 215, 280] and was informed by related efforts in the emerging area of "reproducible research" [40, 106, 238].

[40] In the sense of Claerbout & Karrenbach [58].

After the policy was announced, the editors devoted an issue of the journal to the discussion of the policy and a critique from advisory board member Neils Keiding [77, 141]. Keiding argued that the policy's focus on documentation of data and code ignored the "substantive context" of statistical analysis. In 2010, shortly after the policy was announced, founding editors Zeger and Diggle stepped down [116] and it appears that the policy was no longer enforced after 2011[41]. Ahead of its time in terms of policy and infrastructure, between 2009 and 2011 five articles were evaluated based on the reproducibility workflow. While Peng's earlier work (e.g., [213]) remains influential in the area of computational reproducibility, in later work he advocates for a "preventive approach" to reproducibility based on increased education [161, 209].

Biostatistics is unique among the cases studied in that it is a young journal funded by a charitable trust. Other reproducibility initiatives have been undertaken by mature journals published by larger academic societies.

4.5.1 Initiative Organization

The journal *Biostatistics* was created in 2000 by the Biometrika Trust[42] in cooperation with Oxford Journals and founding editors Scott Zeger and Peter Diggle. According to the journal website:

> *Biostatistics* publishes papers that develop innovative statistical methods with applications to the understanding of human health and disease, including basic biomedical sciences. Papers should focus on methods and applications. Introduction of original methodology should be grounded in substantive problems; there is the opportunity to present extensive analyses of data on the journal's website as supplementary material. Authors are strongly encouraged to submit code supporting their publications. Authors should submit a link to a Github repository and to a specific example of the code on a code archiving service such as Figshare or Zenodo.

[41] The last artifact given the "R" designation was in 2011.

[42] A charitable trust established at the death of Karl Pearson in 1936 to continue ownership of the journal *Biometrika* [64]

Cox [64] notes some of the characteristics of publishing under a charitable trust:

> The ownership of the journal by a charitable trust has massive advantages over commercial ownership. As contrasted with ownership by a society, there are, however, the disadvantages of not having loyalty to a society as the basis for largely voluntary work by referees and others, and also of not having an almost captive subscription base. There are also disadvantages in running what in some respects is a small international business with minimal resources...A final advantage of control by a trust is that the editor is not in any sense a representative of a society and therefore rather more free to exercise judgement, unconstrained by formal guidelines, and to set a distinctive style for the journal.

Oxford Journals is a division of the Oxford University Press, one of the largest academic publishers. Oxford Journals "publishes over 300 journals in the humanities, social sciences, law, science, and medicine, two-thirds of which are published in partnership with learned and professional societies" and has the highest percentage of journals in the top 10% by JCR Impact Factor among publishers with over 100 journals in the ranking[43]. Oxford provides many capabilities used by the journal including online publishing and supplemental data facilities[44].

4.5.2 Initiative History

Historical antecedents to the *Biostatistics* initiative include the earlier work in "reproducible research" [40, 106, 238] as well as Peng and Zeger's work on reproducibility in epidemiology [210, 214, 215, 280]. As a field, epidemiology had gained a reputation for irreproducibility [36]. Zeger, in an editorial for the *Journal of the Royal Statistical Society. Series A*, discusses the importance of transparency in statistics research, arguing that as hypotheses, datasets, and analytical methods grow more complex, so does the risk of incorrect results [280]. In the absence of independent replication, he continues, "transparency of the methods that are employed is essential" and that"[f]or reproducibility to become a research standard,

[43] https://academic.oup.com/journals/pages/about_us

[44] https://academic.oup.com/biostatistics/pages/supp_data

statisticians and their societies must advocate for its implementation. As a first step, our own journals can usefully become reproducible."

Inspired by their work, in 2007 the *Annals of Internal Medicine* announced a policy requiring authors to "include a statement that indicates whether the study protocol, data, or statistical code is available to readers and under what terms authors will share this information" [160]. *Biostatistics* announced its own policy in 2009 [76, 211].

The policy received criticism from the community, including advisory board member Niels Keiding [141]. The editors invited Keiding along with other researchers from statistics and medical communities to comment, published as *Biostatistics* 11(3). Keiding raised concerns about the policy's focus on documentation of data and code, ignoring the "substantive context" of statistical analysis.

> [T]he actual mechanical SAS- (or R-) number crunching of the finally developed master data set is a minor part of the totality of the statistical analysis, and it ridicules our profession to believe that there is a serious check on reproducibility in seeing if somebody else's computer reaches the same conclusion using the same code on the same data set as the original statistician's computer did.

In any effort toward reproducibility or reanalysis, he suggests that "there at least has to be sufficient information to make it realistic for another interdisciplinary group of researchers to understand the substantive context and the strengths and weaknesses of the data."

Breslow [37] worries that such a "seal of reproducibility" would provide a false sense of security, taking attention away from "more serious problems, including selection bias, measurement error, uncontrolled confounding, and small sample size, that affect so much of today's epidemiologic research." Cox and Donnelly [65] considered it a "not merely unnecessary but a misuse of relatively scarce expertise." Quoting Cox and Donnelly in his response, Keiding [142] emphasizes the terminological problem in using "reproducibility":

> In summary therefore, the proposals for *Biostatistics* are to be welcomed even though their name and objectives are misformulated.

After ten years, in 2010 founding editors Diggle and Zeger stepped down and the following editors Molenberghs and Tsiatis appear to have not continued the policy. The AER only reviewed five papers, the last of which was published on November 30th, 2009 ([172]). In 2011, *Science* published a special issue on "Data Replication & Reproducibility" [139] including a perspective piece by Peng entitled "Reproducible Research in Computational Science" [213], perhaps one of the top-cited papers in the area of computational reproducibility.

In later publications, Peng moves away from computational reproducibility, advocating instead for statistical education [209]. Citing the Anil Potti and Reinhart & Rogoff cases, he concludes that improving the quality of science and data analysis requires more than computational reproducibility:

> If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in the form of screening and rejection of poor data analyses by journal referees, editors, and other scientists in the community. Once a poor data analysis is discovered, it can be 'treated' in various ways.

He continues by describing the "medication" approach where peer reviewers and editors make a diagnosis and the "preventative" approach through improved education.

> In much the same way that the epidemiologist John Snow helped end a London cholera epidemic by convincing officials to remove the handle of an infected water pump, we have an opportunity to attack the crisis of scientific reproducibility at its source. Dramatic increases in data science education, coupled with robust evidence-based data analysis practices, have the potential to prevent problems with reproducibility and replication before they can cause permanent damage to the credibility of science.

In 2016, Dimitris Rizopoulos and Jeff Leek took over as co-editors of the journal and announced several initiatives related to the journal, none of which explicitly discontinued nor continued the AER role. Instead they focused on launching a Twitter account, a blog, adding Altmetrics widgets to the journal webpage, simplifying the submission process through direct submission via Overleaf and removing requirements for formatting initial submissions. They even created a "Shiny app" in R to allow authors to explore review times. They revised the data sharing policy to encourage use of Github, Figshare, and Zenodo.

In their analysis of badges for data and code sharing, Rowhani-Farid and Barnett [233] conclude:

> Badges did not appear to have an effect on code sharing as the prevalence ratio was 1.1. When the now broken links were assumed to indicate code sharing, the badge effect on code changed slightly from 0.61% to -2%. This is an unexpected outcome as code is of great importance in the field of biostatistics. A possible explanation behind the lack of badge effect on code sharing could be our definition of code sharing, which might seem traditional compared with the reproducibility policy at *Biostatistics* .

Another explanation is that the policy was not encouraged almost immediately after it was announced and was not maintained by the following editorial teams.

4.5.3 Policies and Guidelines

The *Biostatistics* reproducible research policy and guidelines described in journal's "Information for Authors" [45] and Peng's 2009 editorial [211]. The current policy states:

> **Reproducible Research** Our reproducible research policy is for papers in the journal to be kite-marked D if the data on which they are based are freely available, C if the authors' code is freely available, and R if both data and code are available, and our Associate Editor for Reproducibility is able to use these to reproduce the results in the paper. Data and code are published electronically on the journal's website as Supplementary Materials.
>
> **Code Availability** Authors are strongly encouraged to submit code supporting their publications. Authors should submit a link to a Github repository and to a specific example of the code on a code archiving service such as Figshare or Zenodo.

The "Code Availability" statement was added in 2016.

[45] https://academic.oup.com/biostatistics/pages/General_Instructions

4.5.4 Technical Infrastructure

The primary technical infrastructure is provided by Oxford Journals including the journal website and built-in facilities for supplemental data. The journal website hosts the author guidelines and policy description. Oxford uses ScholarOne for manuscript submission and peer review. Use of Github, Figshare, and Zenodo were adopted in 2016. Also beginning in 2016, editors began to use social media tools including Medium[46] and Twitter[47] to communicate new policies and initiatives.

4.5.5 Artifacts, Identifiers, Badges, and Metadata

Biostatistics was one of the first journals to adopt a "badging" strategy for submitted manuscripts, referred to in policy documentation and related communications as "kite marks."[48] As part of the publication process, a boxed letter or set of letters would be embedded in the PDF or print article based on the policy.

Based on the current policy and workflow, peer review information is available in ScholarOne and the resulting paper is published by Oxford with the embedded kitemark. The reviewed artifacts were under the 2009 policy and artifacts were not necessarily published online as either supplemental information or in a research data repository.

4.5.6 Metrics

Biostatistics does not publish public editor reports, so conventional metrics of impact factor, rank, and submission rates are assumed. In 2016, the editors published a "Shiny" application[49] to allow readers to explore review timelines. Over the course of the policy, only 5

[46] https://medium.com/@biostatistics

[47] https://twitter.com/biostatistics?lang=en

[48] As discussed by Ware [270], "Kitemark" is a registered trademark of the British Standard Institution (BSI) used for product certification. The Kitemark is used primarily to certify that manufacturing products conform to BSI engineering standards of safety and quality. The phrase "kite mark" has been informally used to describe other certification processes, such as the quality of online medical information[73, 88].

[49] `https://jhubiostatistics.shinyapps.io/Biostatistics_Review_Times/`

papers received the “R” designation for reproducible, with the last in 2011. There is no easy way to identify the kite-marking of papers outside of visually scanning the print or online publications.

oo

4.6 Information Systems (IS)

In 2016, the journal *Information Systems* (IS) announced a new “invited reproducibility paper” initiative and introduced a new “Reproducibility Editor” position via an editorial in the journal [55] and press-release from Elsevier [152]. A stated goal of the initiative is to “increase the practice of reproducibility in computational sciences.” Through this initiative, the editors of the journal invite authors of a previously published paper to collaborate on a “reproducibility paper” describing the software, data, and steps to reproduce and extend published results. Reviewers from the community are recruited to validate the claims in the paper and become co-authors on the resulting reproducibility paper. The initiative is described as “the latest effort in Elsevier’s history of exploring the potential for reproducibility in scientific publications.”

The roots of the initiative can be traced back to the “Repeatability and Workability” experiments in the ACM SIGMOD community from 2008-2012 [34, 173, 174]. Dennis Shasha, a co-editor-in-chief of *IS* since 1994, was chair of the SIGMOD 2008 conference and serves on the advisory committee of the current SIGMOD reproducibility initiative. Shasha and his collaborators have been involved in the development of tools to improve the reproducibility of computational research including Vistrails [97] and ReproZip [56].

Of the initiatives studied, *IS* is the only one to require extension as opposed to simple assessment or strict reproduction of results. The *IS* initiative is also unique among the studied cases in that it is supported by a major commercial publisher using their infrastructure, as many other initiatives are supported by academic societies.

4.6.1 Initiative Organization

The *IS* journal was first published in 1975 by Pergamon Press and became closely associated with leading European database conferences including the annual Conference on Advanced Information Systems Engineering (CAiSE) and the International Conference on Extending Data Base Technology (EDBT) [137, 138]. Pergamon was sold to Elsevier in 1991.

Elsevier, part of the RELX group, is one of the world's largest scholarly publishers, accounting for 18% of global research in 2018[50]. Elsevier acquired Aries Systems, developer of the Editorial Manager software commonly used for peer-review and production tracking of journals. Elsevier acquired Mendeley in 2013, but did not fully integrate it into the platform until 2018 [51]. In 2015, Mendeley announced its new "Mendeley Data" service[52].

Dennis Shasha became co-editor-in-chief in 1994 [241]. In 2016, *IS* announced the new Reproducibility Editor position, currently held by Fernando Chirigati, developer of ReproZip reproducibility tool [56]. The current scope of the journal is[53]:

> Subject areas include data management issues as presented in the principal international database conferences (e.g., ACM SIGMOD/PODS, VLDB, ICDE and ICDT/EDBT) as well as data-related issues from the fields of data mining/machine learning, information retrieval coordinated with structured data, internet and cloud data management, business process management, web semantics, visual and audio information systems, scientific computing, and data science. Implementation papers having to do with massively parallel data management, fault tolerance in practice, and special purpose hardware for data-intensive systems are also welcome.

[50] https://www.relx.com/~/media/Files/R/RELX-Group/documents/reports/annual-reports/2018-annual-report.pdf

[51] https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-launches-mendeley-data-to-manage-entire-lifecycle-of-research-data

[52] https://blog.mendeley.com/2015/11/09/put-your-research-data-online-with-mendeley-data/

[53] https://www.sciencedirect.com/journal/information-systems/about/aims-and-scope

4.6.2 Initiative History

The history of the *IS* initiative can be traced back to the "Repeatability and Workability" experiments in the SIGMOD community from 2008-2012 [34, 173, 174]. The SIGMOD Repeatability experiment was an initiative within the SIGMOD conference to introduce repeatability and workability testing into the conference program. *IS* editor Shasha was the Program Committee chair in 2008 and he and collaborators have been involved in multiple community reproducibility initiatives, including the most recent `db-reproducibility.org`. Organizers enumerate benefits to the community including:

1. full specification of algorithms, code, and data helps keep track of the factors that influence the experimental results. Repeatability is thus a way to ensure that there are no hidden factors that influence the results (e.g. compiler settings).
2. a repeatability tester can easily change data, thus testing software in new settings preparing code and data for repeatability leads, without much additional work, to preparing the code
3. for archiving and distribution, thus allowing future researchers to compare their implementations with previous ones.

The initiative was refined in 2009 and 2010 [34, 173] as evaluation was conducted on only accepted papers and reviewer assignment based on hardware/software requirements. Although participation varied over time, the organizers and community anticipated "coming repositories of reproducible experiments" [94]. The initiative was apparently stopped after a later mandatory policy negatively impacted submission rates[54].

The *IS* "invited reproducibility paper" in part reflects the experience of the community in conducting reproducibility evaluations. The invited paper approach provides the editorial team with control over the process, including scaling up over time.

[54] Private communication.

4.6.3 Policies and Guidelines

In addition to the original editorial [55], information about the initiative is posted under the "Guide for Authors"[55] on the journal website in the section "Invited Reproducibility Paper." Authors are instructed to select the "Reproducibility Section" in Editorial Manager, which is notably no longer an option. Reviewer Guidelines[56] are provided online.

4.6.4 Technical Infrastructure

The *IS* journal infrastructure is based primarily on the Elsevier family of tools. Journal policies and guidelines as well as online access to published papers are provided by Elsevier. Editorial Manager is used for manuscript submission and peer review. All supplementary materials are published to Mendeley Data. Outside of the Elservier tools, the *IS* initiative guidelines encourages the use of Github/GitLab for code hosting and Docker and ReproZip for packaging research environments.

4.6.5 Artifacts, Identifiers, Badges, and Metadata

Based on the current workflow, artifacts from the process include the peer review communication in Editorial Manager, reproducibility paper published by Elsevier, artifacts in Mendeley Data and optionally Github/GitLab.

DOIs are assigned to the original paper, reproducibility paper, and Mendeley Dataset by Elsevier. The original and reproducibility papers are linked via "Refers to" and "Referred to by" citations in the online article. The Mendeley Data record includes the citation for the reproducibility paper.

Reproducibility papers are published in a "Reproducibility Papers" section of the journal, but otherwise receive no badge or additional metadata in Elsevier's system (e.g., in

[55] https://www.elsevier.com/journals/information-systems/0306-4379/guide-for-authors

[56] http://fchirigati.com/files/is/GuidelinesReviewers.txt

accordance with ACM or COS practices).

4.6.6 Initiative Metrics

In 2018, *IS* ranked 82/175 based on JCR 2-year impact factor. Elsevier reports numerous metrics about all of its journals including: CiteScore, Impact Factor, 5-Year Impact Factor, article influence and eigenfactor, SNIP Source Normalized Impact per Paper (SNIP), review speed, online article publication time, and author reach. Mendeley Data reports views and downloads for the published dataset. Github reports stars and forks. As of writing, 3 invited papers have been identified as invited reproducibility papers.

4.7 Journal of the American Statistical Association (JASA-ACS)

In 2016, *JASA-ACS* announced a new policy and process to improve data and code sharing practices and assess the reproducibility of published research. Called a "reproducibility initiative," the policy set "a minimum standard for reproducibility in statistical scientific research" [98]. The journal created a new editorial role – the Associate Editor for Reproducibility (AER) – who would be responsible for reviewing materials described in a checklist (ACC Form) submitted by authors and assess the reproducibility of the work presented, as well as developing policies and implementation strategies for the journal.

The new policy was one of a number of activities undertaken by the American Statistical Association (ASA) in response to external factors from the open science movement and the emerging "crisis" in scientific credibility and reproducibility [16, 59, 134]. Other activities included the publication of a policy statement on the use of p-Values [272]; revisions to the ASA *Ethical Guidelines for Statistical Practice* [11]; work with *Nature* and *Science* on involvement in statistical review of research [271]; and recommendations for funding agencies related to reproducibility [12].

4.7.1 Initiative Organization

The ASA is one of the largest professional associations in the field of statistics and responsible for the publication *JASA*, its flagship journal, along with 17 other publications. *JASA* publishes two different sections – "Application and Case Studies" and "Theory and Methods." The reproducibility initiative applies only to *JASA-ACS*, although there have been recent discussions about expanding the policy to *JASA-TM*[57]. In 2018, *JASA* ranked 5/123 based on JCR 2-year impact factor in the "Statistics and Probability" category.

The ASA Committee on Publications oversees publication policy and includes representatives from each of the 17 ASA publications[58]. In 2014, the ASA convened a Committee on Data Sharing Reproducibility which was responsible for "the issues surrounding data sharing from a statistical perspective."[59]

Beginning in 1970, *JASA* articles were divided into "Applications" and "Theory and Methods" with separate editors [234]. The "Applications" section was renamed to "Applications and Case Studies" in 1987 to "emphasize that careful analysis of data of substantive importance may be published in *JASA* even if they do not have methodological innovations" [239]. According to the author guidelines, *JASA-ACS* currently publishes original articles that present statistically innovative analysis that are scientifically relevant; contribute to a scientific field through the use of statistical methods; present new and useful data; use empirical tests to examine the utility of a statistical technique; and/or evaluate the quality of important data sources.

JASA-ACS has a rotating editorial term. Montserrat Fuentes was Coordinating Editor for *JASA* and Editor of *JASA-ACS* when the policy was first implemented. The journal

[57] Private communication.

[58] Taylor & Francis publishes 9 of the 17 journals.

[59] `https://www.statisticsviews.com/details/news/6165391/ASA-launches-committee-on-Data-Sharing-and-Reproducibility`

is published by the Taylor & Francis Group and uses ScholarOne[60] for manuscript submission and peer review. Since 2014, the Taylor & Francis platform uses Figshare[61] to host all supplemental data, including the ACC Form associated with each article. The *JASA-ACS* initiative began with three AERs in 2016 and has since grown to six. The AERs are responsible for the definition of the review process as well as the assessment of the ACC form and any associated materials. AERs are regular research community members (i.e., not students or practitioners, as in other initiatives).

4.7.2 Initiative History

There are a number of historical antecedents and events related to the *JASA-ACS* initiative. These include early community practices of code sharing and software distribution [128, 129] and the development of techniques and technologies for distributing "reproducible research." [40, 106, 163] The *JASA-ACS* initiative began at a time of increasing concern about credibility and rigor of scientific research and was motivated largely by external events. The focus of the *JASA-ACS* initiative has been on software quality and the transfer of new methods.

As discussed in Chapter 2, the practice of sharing code and software began in the statistics community in the 1960s with the introduction of the "Algorithms" section in RSS *Applied Statistics* [192], modeled after a similar initiative in the ACM [129]. In the late 1980s, with the emergence of techniques for electronically distributing software, CMU established StatLib [153], modeled after Netlib in mathematics [78]. The statistics community has also developed widely-adopted programming environments including S [27] and R [133] and related software distribution methods including CRAN [130]; popular methods for literate programming [163, 232, 278]; and methods for the publication of "reproducible research" [105, 106]. The statistics community developed the concept of the "research compendium" now widely used in other fields [176].

[60] Previously named Manuscript Central, adopted by T&F in 2008 [82]).

[61] Figshare was established in 2011 and partnered with T&F in 2014.

The discussion of reproducibility in the ASA was motivated primarily by external events, such as the Duke scandal [16] as well as failures to reproduce studies in cancer research [28], drug testing [222] and psychology [59, 195]. These same events are cited in the adoption of stricter editorial policies in journals such as *Nature Methods* [6]. Although these events concerned external fields, it is worth noting that it was the statistics community that published many of their findings. *Annals of Applied Statistics* published Baggerly & Coombs original findings [15] and the ASA *Amstat News* published their subsequent editorial on reproducible research [14].

In the 2014 London Workshop, the capstone event of the 175th anniversary of ASA, issues related to reproducibility were considered a priority for the community [171]. This resulted in several initiatives within the ASA, including the Statement on p-Values [272]; updates to the ASA *Ethical Guidelines for Statistical Practice* [11]; work with *Nature* and *Science* on involvement in statistical review of research [271]; recommendations for funding agencies related to reproducibility [12], as well as the *JASA-ACS* reproducibility initiative.

The *JASA-ACS* initiative followed earlier work in the journal *Biostatistics* [211] and the ACC Form was adapted from a LaTeX template used for artifact evaluation in the ACM[62].

4.7.3 Policies and Guidelines

JASA-ACS "Reproducibility Initiative" policy is included in the general "Instructions for Authors"[63] on the journal website hosted by Taylor & Francis. The guidelines include the following statement:

> To enhance the reproducibility of published research, authors submitting to *JASA* Applications and Case Studies will be expected to provide relevant code and data at the time of submission, or to provide a brief explanation why code and data are not available.

[62] `https://ctuning.org/ae/submission_extra.html`

[63] `https://amstat.tandfonline.com/action/authorSubmission?journalCode=uasa20&page=instructions`

The policy description includes a link to the Author Contributions Checklist (ACC) form. The form is intended to document artifacts associated with a manuscript, particularly code, data, and steps required to reproduce findings. The form is the only artifact required by the authors and captures the following information:

1. Data availability, permissions, licensing, link to external repository, provenance of source data, metadata (data dictionary), and version information
2. Code licensing, link to external repository, version information, required software, and hardware requirements
3. Instructions for how to reproduce results presented in the paper
4. Expected runtime of workflow

The AERs also maintain a set of guidelines and evaluation criteria for assessing completeness, which are not public. According to the policy, reviewers assess the potential reproducibility but are not required to run the code and confirm results:

> During the review process, the reviewers and editors will assess the potential for the work to be reproduced based on the code and data provided. However, as with other aspects of a manuscript, it is ultimately the authors' responsibility to ensure that the code and data are of high quality and that the work is reproducible.

4.7.4 Technical Infrastructure

The ASA maintains the `amstat.org` website for communication with ASA membership and the general public. The website hosts the ASA *Ethics Guidelines*, *AMSTATNEWS* and community blog, which are sources of information for this case report. The *JASA-ACS* journal website is hosted by Taylor & Francis which provides author submission guidelines, including the reproducibility policy, and access to online papers and supplemental information in Figshare. Taylor & Francis use ScholarOne, a commercial tool used to manage the manuscript submission and review process. Taylor & Francis use Figshare for all supplemental information submitted by authors. In ScholarOne, the AERs operate under a single fake reviewer account ("Dr. Reviewer for Reproducibility") and reviewer assignment is handled via email. The *JASA-ACS* AER team has established a separate Github organization to host repositories for final verified materials for each article [64]. Aside from Github and email, the*JASA-ACS* does not maintain any custom infrastructure to track the review process. Any computational resources or software licenses are provided by the individual AER's institutions.

4.7.5 Artifacts, Identifiers, Badges, and Metadata

Based on the current policy and workflow, the publication process results in the paper published by Taylor & Francis, ACC Form in Figshare as supplemental material, and final verified materials in Figshare and *JASA-ACS* Github. Peer review information including AER reports are held in ScholarOne.

DOIs are assigned to the paper by Taylor & Francis and the dataset by Figshare (Nature/Springer). The online materials and PDF contain links to supplementary materials in Figshare, which are also displayed with the online record. The Figshare record contains a link to the paper. The Github repository contains the citation of the original paper, but is not referenced in either the online publication or the Figshare record.

The PDF of the paper includes a checkmark indicating that "These materials were reviewed for reproducibility." This does not appear in the online article and it is not possible to filter searches based on this attribute. Article abstracts include the statement "Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement." These are the primary indicators that an article has been reviewed, but no reference to the specific policy or guidelines.

[64] `https://github.com/jasa-acs`

4.7.6 Initiative Metrics

There are no annual reports available and it is unknown what metrics *JASA-ACS* editors consider in the management of the journal. In addition to common journal metrics (e.g., JCR impact factor, JCR rank, Google Scholar h-index), the Taylor & Francis platform provides article views, citations, and Altmetrics; the Figshare platform provides dataset views, downloads, and citations; Github provides forks and stars.

During the interviews, limitations of the editorial platform and ACC form with respect to reporting were mentioned. Desired metrics included number of rejected manuscripts prior to AER review as well as percent of submissions in a given programming language or with public data.

As of February, 2020, there are 52 article repositories in the *JASA-ACS* Github.

4.8 Transactions on Mathematical Software (TOMS)

In 2015, the ACM *Transactions on Mathematical Software* announced a new initiative for the reproduction pf computational results published in the journal called the "Replicated Computational Results" (RCR) review [122]. The initiative established a policy and process for reproducing results presented in research papers comparable to the existing *TOMS* algorithm review. Since it was established in 1975, *TOMS* – one of the oldest journals of the ACM – has published two primary classes of articles: research papers that present original research and algorithms papers that describe particular implementations. For algorithms, the journal adopted a specialized review policy, refereeing process, and system for the dissemination of mathematical software [129]. The new RCR policy is the first time that the journal has adopted similar practices for research articles.

TOMS was one of the earliest initiatives within the ACM community to adopt the nascent badging standards and journal leadership has been involved in related initiatives, including

Supercomputing. The *TOMS* community shares many of the challenges to reproducibility in the HPC community. Although *TOMS* has long published mathematical software associated with algorithms submissions, it is unique among the studied cases in not requiring authors to publish computational research artifacts related to the RCR process.

4.8.1 Initiative Organization

Please see the *Supercomputing* case profile in this chapter for information about the organization of the ACM, Digital Library, and related committees. *TOMS* is one of the oldest journals published by the ACM and is closely associated with *CALGO*, one of the first repositories for research software. *CALGO* and its relationship to*TOMS* is discussed in the next section.

TOMS has a rotating editorial term with Editor-in-Chief, Algorithms Editor and RCR Editor positions. The RCR Editor position was established with the RCR policy in 2015. The RCR Editor is responsible for recruiting reviewers, who are typically peers or expert practitioners.

4.8.2 Initiative History

The roots of the *TOMS* RCR initiative can be found in the origins of the journal itself. *TOMS* was established in 1975 to expand opportunities to publish results based on mathematical software while also replacing the "Algorithms" department of CACM [229]. From the beginning, *TOMS* was associated with *CALGO* [128, 129] and the Algorithm Distribution Service [245]. From the first issue,*TOMS* established a new Algorithms Policy and refereeing process [93]. The focus on software, practice of distributing algorithms, and implementation of a refereeing process established the groundwork for the RCR process that would be adopted 40 years later. Notably, the algorithm policy did not apply to the submission of research papers.

According to Rice [229], *TOMS* accepted two broad types of submissions: 1) fundamental research papers on the analysis and critical evaluation of computer programs; and 2) practically oriented, concrete research and development in areas including linear algebra, polynomial manipulation, and non-linear programming. These later became "research" and "algorithm" submissions.

The Algorithms Policy defines the standards of publication for algorithms in TOMS. Again according to Rice [229]:

> TOMS will raise the requirements in other aspects of computer programs. This includes insistence on excellent documentation, good use of modularity and structure in the program, adherence to language standards, and thorough testing (and evidence thereof) of the program by the author.

The first Algorithm Policy was defined by Lloyd Fosdick [93], the former Algorithms Editor for the CACM. The policy defined allowable languages (initially Fortran, Algol, and PL/1) and criteria for "substantial contribution." It also established requirements for documentation, copyright and testing. Under the new refereeing process:

> Evidence of reasonable testing with the test environment described. Before publication the program will be independently compiled and executed, and some of the tests supplied by the author will be repeated.

CALGO is a collection of algorithms and software published by ACM journals beginning with the *Communications of the ACM* (CACM) in the 1960s. *CALGO* became part of *TOMS* when it was established in 1975. Submitted algorithms underwent a referee process for "originality, accuracy, robustness, completeness, portability, and lasting value." Hopkins [128, 129] details the history, including the steps required to maintain and "renovate" the *CALGO* library in the late 1990s. This required changes to previously submitted software to improve organization, coding practices, portability, and construction of test drivers in cases where they were missing. *CALGO* is viewed as a library of reusable mathematical software, similar to many commercial offerings and includes many popular packages such as BLAS, LAPACK, and LINPACK.

Hopkins [129] argues that publishing algorithms in *CALGO* fosters research by providing immediate access to state-of-the-art software. The referee process provides certification of the software and added status over self-publication. The notes that, although the process is rigorous and requires extensive testing, there is no way to unpublish the software if found to be defective or superseded by better approaches, although warnings are added to the comments.

Until the 2015 RCR policy, *TOMS* did not include a review process for research articles. Crowder, Dembo, and Mulvey [66] provide an early critique of standards for computational experiments in *TOMS*. They note the articles involving the use or development of mathematical algorithms require a high-degree of mathematical expertise by the authors, but the same didn't apply for computational experiments. They claimed that journals were less concerned with scientific design and reporting than leading journals in the social sciences. "Very rarely can a published computational test of a mathematical algorithm be completely reproduced – something which is a basic criterion in scientific research. Even worse, important parameters of the experiment are often not even reported."

In 2015, then Editor-in-Chief Heroux introduced the new RCR policy and process [122]. He notes that the RCR process is intended to supplement the algorithms review process, adding the replicability expectation. "The RCR policy as stated here should be clear for research papers, which typically do not include a review of the discussed software, nor the software itself." This editorial introduces the new Associate Editor for RCR position, citing the ICERM [252] and Vancouver [166] meetings as motivations for the policy. In a separate report, Heroux [123] argues that expectations for reproducible computational results requires higher quality software processes, documentation, source code management, etc. These all have an impact on programmer productivity and software sustainability.

While not directly related to the RCR initiative, key figures in *TOMS* history were also involved in developments related to Problem Solving Environments (PSEs) in the 1990s and 2000s that are part of the history of reproducible research [230].

4.8.3 Policies and Guidelines

The ACM maintains the `https://acm.org` website and the ACM Digital Library used by *TOMS*. The RCR policy is documented on the journal website[65]. The policy was announced in an editorial [122].

4.8.4 Technical Infrastructure

TOMS uses the ScholarOne system for manuscript submission and papers are published via the ACM Digital Library. Computational resources used during RCR review are provided by the author or reviewer. No other infrastructure is involved in the initiative. While algorithms submissions are published via *CALGO*, artifacts from research articles are not required to be published.

4.8.5 Artifacts, Identifiers, Badges, and Metadata

The primary artifact of the *TOMS* RCR process is the RCR report, published in the ACM Digital Library. In 2015, along with the RCR policy, *TOMS* adopted the nascent ACM badging standards. Between 2015 and 2019, *TOMS* retroactively applied the "Artifacts available" and "Artifacts Evaluated and Reusable" badges to the nearly 500 algorithm papers published since 1975. Three papers have received the RCR designation and "Results Replicated" badge since the policy was established in 2015.

[65] `https://dl.acm.org/journal/toms/replicated-computational-results`

4.8.6 Initiative Metrics

No specific metrics are mentioned for the*TOMS* RCR initiative. Donoho & Stodden [79] report on the number of lines of code submitted to *TOMS* between 1960-2012, observing exponential growth. *CALGO* (`https://calgo.acm.org` lists software associated with 428 papers. As of February 2020, the ACM Digital Library contains 494 *TOMS* papers with the "Artifacts Available" and "Artifacts Evaluated and Reusable" badges along with 3 with the "Results Replicated" badge based on the RCR review process.

4.9 Conclusion

This chapter has presented the summary case profiles for the two primary and five supplemental cases used in this study. These case profiles, along with information presented in the Appendices, serve as the basis for the multiple-case analysis used to develop the results presented in the following chapters.

Chapter 5

Verifying computational transparency and reproducibility

> *[T]he only way to understand and evaluate an empirical analysis fully is to know the exact process by which the data were generated and the analysis produced.*
>
> – King, 1995

5.1 Introduction

The initiatives at the center of this study are broadly concerned with improvements to the quality of reported research in their respective fields through the adoption of policies to assess the transparency and reproducibility of computational results. In response to both factors internal and external to their research communities, they have implemented changes to the peer review process, established new editorial roles, increased the information required of authors and the effort required of reviewers to ensure adherence to new standards of publication. They have adapted to gaps in editorial and publishing infrastructure, often bearing the cost of innovation.

In this chapter I investigate the question (RQ1) *how are computational transparency and computational reproducibility operationalized through publication reproducibility audits?* In the terms of Radder, reproducibility *of what* and *by whom*? I also address the related question of what led these communities to propose and ultimately implement sometimes intensive new requirements on scientific communication. I develop a set of characteristics of the initiatives as they relate to the mechanics of review and verification to better understand the similarities and differences observed between initiative policies and workflows. I compare the metrics used to assess initiative impact, infrastructure used and the key challenges identified by study participants.

This chapter is organized as follows. In the next section, I briefly summarize the methods used to address the research questions followed by a summary of related work. I present the results of the qualitative analysis, including a detailed look at the motivations and characteristics of each initiative; metrics used to evaluate impact and effectiveness; key infrastructure components as well as challenges reported by study participants. This is followed by a discussion of key findings.

5.2 Methods

The results reported in this chapter are based on the qualitative coding and analysis of interview transcripts and documentary evidence. The interview protocol, documentary evidence, codebooks, and detailed workflows implemented by each initiative are provided in Appendices B, C, D, and E respectively. All qualitative analysis was conducted using ATLAS.ti (v8.4.4).

5.3 Related Work

Efforts to assess the transparency and reproducibility of the results of computational research are not new. For decades, researchers within several communities have undertaken systematic studies of the reproducibility or replicability of previously published findings. Related earlier efforts can also be found in policies for the review of software and algorithms as well as data and code sharing.

Over the years, a number of journals and conferences have explored the adoption of policies to encourage and enforce data and code availability requirements. This includes early efforts in economics [10, 74, 170, 30], political science [2, 87, 184, 276], and computer science [156, 173, 174]. Many of these are antecedents to the initiatives that are the focus of the current study.

Also related are policies and workflows for the review of algorithms and software. For example, ACM *TOMS* has had dedicated algorithm peer review since its inception in the mid-1970s [129]. Similar policies were adopted by *Applied Statistics* [4, 5] and can more recently be seen in the peer review policies of journals such as the Journal of Open Source Software (JOSS) or SoftwareX. Algorithm and software review, while related, has typically not involved confirmation of the reproducibility of specific scientific results or findings based on their use.

Over the past four decades, a number of studies have reported the results of efforts to assess the reproducibility or replicability of findings from computational research. Studies have been undertaken in economics [32, 47, 74, 102, 179, 181, 182], parallel computing [42, 127], systems research [61], and computational physics [254, 255]. These studies typically involve a sample of published research from one or more venues and either the assessment of potential reproducibility through availability of certain types of information or attempted reproduction of computational results.

Several studies have looked at data and code availability policies. As detailed in Chapter 2, the *JMCB* study explored the effect of policies changes on the availability of materials and reproducibility of empirical economics research [74]. Stodden et al. [253] investigated data and code availability policies across 170 journals as they relate to journal rankings and impact. Key [143] explores the affect of policies on materials availability in political science journals.

	AEA	**AJPS**	**Biostatistics**	**JASA-ACS**	**IS**	**SC**	**TOMS**
Discipline	Economics	Political Science	Statistics	Statistics	Computer Science	Computer Science	Mathematics
Name	Data and Code Availability Policy	Verification Policy	Reproducible Research	Reproducibility Initiative	Invited Reproducibility Paper	Reproducibility Initiative	Replicated Computational Results
Dates	2019 - [a]	2015 - [b]	2009 - 2011[c]	2016 -	2016 -	2015 -	2015 - [d]
Role	Data Editor	Multiple	AER	AER	AER	Reproducibility Chair	RCR Reviewer
Incentive	Monetary	Monetary	Position	Position	Paper	Volunteer	Paper
Mandate	Mandatory	Mandatory	Opt-in	Mandatory	Invited	Mandatory	Opt-in
Range	Partial reproduction	Full reproduction	Full reproduction	Materials only	Full reproduction with extension	Materials only	Full reproduction
Who	Student	Student, Practitioner	AER	AER	Peer	Peer	Practitioner
Blindness	Single open	Single open	Double open	Single open	Double open	Double-open	Double open
Parent	AEA (association)	MPSA (association)	Biometrika Trust (trust)	ASA (association)	Elsevier (private)	ACM/IEEE (association)	ACM (association)
Type	Multiple journal	Single journal	Single journal	Single journal	Single journal	Conference	Single journal
Publisher	AEA	Wiley	Oxford	Taylor & Francis	Elsevier	ACM	ACM
Repository	OpenICPSR [e]	Dataverse	Figshare/Zenodo [f]	Figshare, Dataverse, Github	Mendeley Data	None	None
Editorial	ScholarOne	Editorial Manager	ScholarOne	ScholarOne	Editorial Manager	Linklinks	ScholarOne
Artifacts	> 200	> 200	< 5	> 50	< 5	> 50	< 5

[a] AER implemented its first Data Availability policy in 1986

[b] *AJPS* implemented its first Replication policy in 1994

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

[d] *TOMS* implemented its Algorithm Policy in 1975

[e] AEA supports multiple repositories, but OpenICPSR is the primary

[f] *Biostatistics* had no recommended repository until 2016

Table 5.1: Summary of reproducibility initiative characteristics (as of February 2020)

5.4 Results

This section reports the results of the cross-case comparison of the seven reproducibility initiatives with respect to the questions of initiative operationalizations and motivations. The comparison identified several initiative characteristics that influence operational workflows. This includes who is responsible for the review process, the expertise required of the reviewer, reviewer incentives, policy mandates, as well as organizational factors such as the parent organization or publisher. The results of this comparison are summarized in Table 5.1. In this section, I also present a comparison of the infrastructure used for each initiative as well as a summary of additional challenges reported by interview participants.

5.4.1 Motivations

The question of initiative motivation was addressed through a combination of key informant interviews and analysis of documentary evidence. Key informants reported a variety of motivations for the initiatives, ranging from broad appeals to the improvement of research quality; responses to external and internal events; concerns about research carelessness; and software quality, trustworthiness, and reusability.

One editor discussed concerns about the ethics of publishing quantitative results in a field where the technical complexity of research is increasing:

> I think generationally, we were not trained adequately as scientists in using quantitative data, on how to document these details, and that on top of the huge shifts in software and analytical techniques and complexity, just really started to add up...[T]hat a leading journal would publish something, and then scholars say, "I really don't know where that data is" just seems ethically wrong and professionally wrong. [1-8]

Another editor reported a broader sense within their community that publication of computational research artifacts should be a matter of practice, independent of other factors:

> So the whole point of the process is to improve science. That's a fuzzy, but laudable goal. And whether or not it will actually do so, will we actually see something going on with articles that are replicated once they're out there? Is this actually a competitive advantage? Most of my colleagues at other [journals] that I've talked to are more along the lines of this is just what we should be doing, period. [3-1]

In a statistics journal, one editor reported the motivation behind the initiative being less about potential errors in published research and more about improving the quality of associated software to facilitate reuse. They noted:

> I think it hasn't been so much about checking that people are not trying to put bad results. It's more about making sure that the results are of high quality and then that we can then take forward those tools and bring them out to the community in the future. [3-2]

A similar motivation was expressed in mathematics, where one editor noted that software quality was a primary concern for the initiative:

> Fundamentally, that was what I was worried about, as a community, we're producing software that's supposed to be used in high consequence environments. We weren't putting in the effort needed to make it trustworthy. That is a fundamental issue. I had a very specific reason why to pursue reproducibility because I wanted to make sure that the software products that our communities were working on we're up to the task of making science credible. [4-3]

From these responses, we see several different broad areas of concern motivating the initiatives. These include professional ethics, potential errors or carelessness, as well as the quality and reusability of published software.

5.4.2 Characteristics

Table 5.1 summarizes key characteristics of each reproducibility initiative (as of February 2020). Table 5.2 summarizes the observed characteristics. In this section, I present a detailed comparison of these characteristics across the different initiatives.

Characteristic	Examples
Role	Associate Editor for Reproducibility, Reproducibility Editor, Data Editor, Curator, Verifier
Expertise	students, expert practitioners, peers
Incentives	position, publication, financial compensation, voluntary
Mandatoriness	mandatory, opt-in, invited
Blindness	Single open, double open
Range	materials only, partial reproduction, full reproduction, full reproduction with extension
Materials availability	review only, general repository, specific repository

Table 5.2: Summary of observed initiative characteristics

Roles

In each of the seven initiatives, new roles were created with specific responsibilities for the reproducibility assessment process. In two cases (*Biostatistics*, *JASA-ACS*), a new associate editor role was created for the journal with responsibility for the entire assessment process, the Associate Editor for Reproducibility (AER). This includes the review of materials submitted with each manuscript, including code re-execution and results evaluation. In three cases (*IS*, *TOMS*, *SC*), a new role was created with responsibility for recruiting reviewers to participate in the assessment process, the Reproducibility Editor or Chair. The recruited reviewers are responsible for the actual review of materials, at varying depths. In one case (AEA), a new Data Editor role was created association-wide. This editor manages the assessment process for materials submitted to multiple journals through the recruitment and training of undergraduate and graduate students responsible for the actual review process. In one case (*AJPS*) staff at a data archive are similarly responsible for the management of the assessment process. Professional curators handle the review of submitted materials and recruit and manage graduate students and staff statisticians who are responsible for code re-execution and results evaluation. Since the archive staff are not part of the editorial team, the managing editor serves as liaison between authors, editors, and archive staff.

The number of individuals in these roles, as expected, depends on the number of manuscripts being processed. While *Biostatistics* had a single AER, *JASA-ACS* currently has six. *IS* and *TOMS* each have a single editor who has recruited reviewers for fewer than five papers. The *AJPS* currently has two curators and six student verifiers while AEA has one editor and ten student verifiers. For the AD/AE review, *SC* 19 had a single chair with eight reviewers.

Reviewer Expertise and Incentives

These initiatives also reflect different approaches to reviewer expertise and incentives. In two cases (*Biostatistics*, *JASA-ACS*), the review is conducted by an associate editor, a peer and expert in the field, where the incentive to conduct the review is the editorial role itself. In one case (*IS*), multiple reviewers are recruited from the community, also peers and experts, to participate in the reproduction. The incentive in this case is authorship on the resulting reproducibility paper. In one case (*SC*), peer reviewers are also recruited from the community, with no additional incentive. In one case (*TOMS*), recruited reviewers are typically expert practitioners, selected because they represent potential users of the submitted research software. The incentive in this case is the published RCR report. In one case (*AJPS*) reviewers (or verifiers) are advanced doctoral students or professional statistician who are paid and gain potentially valuable experience. Similarly, in the AEA initiatives, reviewers (or verifiers) are trained undergraduate or graduate students who are paid.

These results suggest three broad levels of expertise: students, expert practitioners, and peers; as well as four incentive structures: editor position (*Biostatistics*, *JASA-ACS*), publication (*IS*, *TOMS*), financial compensation (AEA, *AJPS*), and voluntary (*SC*). The SCC RC is unusual since the results are reproduced by committee members (editor position) and also students (voluntary).

Mandatoriness

The seven initiatives have three distinct requirements for participation: mandatory, opt-in, and invited. Four initiatives (AEA, *AJPS*, *JASA-ACS*, *SC*) now have mandatory submission requirements. The depth of review ranges from appendix assessment (*SC*) to full reproduction (*AJPS*). Two initiatives[1] (*Biostatistics*, *TOMS*) require authors to voluntarily opt-in to the reproducibility review process and both conduct full reproductions. One initiative (*IS*) requires an invitation from the journal's editors to participate. Mandatoriness is associated with the number of artifacts reviewed. Those with mandatory policies have much higher participation rates than those without. *Biostatistics*, *TOMS*, and *IS* have so far reproduced fewer than 5 papers each over a period of five or more years. The SCC RC can also be seen as an invited paper.

Range

The seven initiatives represent four ranges of reproducibility assessment. *Materials only* requires that reviewers only assess author provided materials without any attempt at reproduction (i.e., running the code) (*JASA-ACS*, *SC*). *Partial reproduction* occurs when reviewers may optionally reproduce a subset of results. This is reflected, for example, in the AEA policy statement that code will be re-executed "when feasible" (AEA). *Full reproduction* occurs when reviewers are required to re-execute all code and assess results as compared to the published manuscript (*AJPS*, *Biostatistics*, *TOMS*). *Full reproduction with extension* includes full reproduction and requirements that reviewers attempt to extend the submitted work, for example through changes to parameters, input data, input conditions, etc (*IS*). The SCC RC can be seen as a partial reproduction with extension, since are only reproducing a subset of results in the paper.

Blindness

The initiatives take two different approaches to the anonymity of authors and reviewers to each other, or blindness. Reviews are either single-open, where the reviewer is unknown to the author but the author is known to the reviewer, or double-open, where both the author and reviewer are known.

[1] *SC* was opt-in for 2015-2018

One editor involved with a single-open review explained:

> I think it is in keeping with the transparency. I think it forces the author to do a better job explaining what they did. If they could just shoot it all over to an email or pick up the phone and call someone and sort of convince them that this is what they did and it should be released, then they're not doing the hard work of making things more transparent and accessible for others. [1-1]

In an initiative that relies on student verifiers, the single-blind review is intended in part to protect their identity. A student verifier discussed their appreciation of this approach:

> So I think it probably is not the most efficient it could be, but I think it makes sense sort of from selfish professional perspective because I can definitely see some senior scholars not looking favorably on 26 year old grad students who are holding up their big publication and might hold that against them in the future.

In several of the cases, the reproducibility review is viewed as a supportive process that lends itself to double-open review. Two editors described their views on double-open review:

> So the authors and the reviewers knew each other and the reviewer... In fact calling it 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 [submission] 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. [4-7]

> [T]he authors know who the reviewers are because we want to encourage this interaction between the authors and the reviewers. Sometimes we are working with research. We do our own code. It works on our computer, and perhaps when we create a bash script to install something, and then it won't work in a specific machine for some reason. Are the authors to be blamed? Probably not. It was just a mistake, or maybe it was a typo. Maybe they didn't know that this could happen. [4-2]

Materials availability

Materials availability refers to how initiatives encourage or require authors to make their materials available, outside of the published manuscript. This is both for the pre-publication review process as well as post-publication access. There are three broad classes of availability: author responsibility, supplemental information and archival deposit. Increasingly, publishers are moving away from traditional supplemental information strategies and using integrated data archive infrastructure. Note, most initiatives have special considerations for proprietary and protected information, which is covered later.

Two initiatives (*AJPS*, *IS*) require authors to deposit materials in a specific archival repository (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 archival repositories (e.g., Zenodo and Figshare). One initiative (*JASA-ACS*) requires submission of supplemental information, which is made available via Figshare and Github. One initiative (*TOMS*) only requires that authors make materials available during the review process and offers multiple different approaches including guest access to remote systems.

5.4.3 Metrics

When considering implementing new policies or initiatives, associations, publishers, journal editors and conference organizers often turn to common operational metrics such as the number of submissions and revisions or decision and processing times. These measures are used to assess the effect of policy or operational changes. Impact measures, citation and circulation rates are commonly used to represent the quality and reach of a journal. Other measures of reuse, while less common, are available from data repository infrastructure including the number of views or downloads of a dataset as well as data citations.

Table 5.3 summarizes the metrics available to initiatives identified during case analysis. These include impact measures, operational measures, measures of re-use, and measures of the reproduction process. A few of metrics have been used by the different initiatives to measure the effect of new policies. Two cases report monitoring JCR impact factor (*AJPS*, Biostatistics). Three cases report monitoring submission and acceptance rates (*AJPS*, *JASA*-

Metric type	**Description**
Impact	
Impact factors	(e.g., JCR 2- and 5-year, Google h-index, SNIP)
Citations	Number of citations per paper
Circulation rates	Historic measure of journal reach
Operations	
Submissions	Journal/conference submission and acceptance rates
Revisions	Number of revisions
Decision times	Journal/conference decision times
Processing times	Journal/conference processing times
Policy conformance	Number of papers that conform to a policy (e.g., AER)
Re-use	
Views	Number of times a publication, file or dataset is viewed
Downloads	Number of times a publication, file or datasets is downloaded
Data citations	Number of publications citing a particular dataset/version
Forks/stars	Number of times a Github repository has been "forked" or "starred"
Reproduction	
Curation time	Amount of time in curatorial review [135]
Verification time	Amount of time required for verification [135]
Replications	Number of replications of a study [117]
Errors	Number and types of errors [117, 2]
Cost	Cost of curation/verification [87]

Table 5.3: Initiative metric types and examples

ACS, *SC*). *AJPS* additionally reports publication delay caused by the verification process and captures information about the number of resubmissions, errors found, and cost. Metrics listed under the "Re-use" section are available from existing infrastructure (e.g., research data repositories and version control systems) and are not designed the initiatives, but may be included in reporting.

5.4.4 Infrastructure

Each of the studied initiatives is part of an existing peer review and publishing process, relying on existing communications, editorial, and publishing infrastructure. This section reviews the different types of technical infrastructure required to support these initiatives, as summarized in Table 5.4. This includes infrastructure for communicating policies; publishing and editorial management; data and code sharing; reproduction and verification; as well as tools used by individual researchers.

Type	**Description**
Communication	
Website	Association, journal or conference website used to communicate policy and guidelines to authors
Social media	Blog and microblog services used to informally communicate information related to policy and guidelines
Publishing/editorial	
Digital library	Association or publisher library of published papers (e.g., ACM DL, Scopus, Wiley Online)
Editorial management	Software used for submission and peer review process (e.g., Editorial Manager, ScholarOne, Linklinks, HotCRP)
Data/code	
Archival repository	Data and code repositories intended for long-term preservation of published research artifacts (e.g., Dataverse, OpenICPSR, Mendeley Data, Zenodo, Figshare)
Software distribution	Software-specific distribution (e.g., pip, CRAN, Github)
Reproduction/verification	
Workflow management	Custom workflow management tools required to track reproduction and verification process (e.g., JIRA, Odum Dashboard, custom databases)
Intermediate storage	Filesystem or other storage to track intermediate submissions during the verification process (e.g., Bitbucket, Odum filesystem)
Compute infrastructure	Computational hardware required to re-execute submitted code(e.g., cloud VMs, batch compute systems)
Software licenses	Licenses for proprietary software (e.g., MATLAB, STATA, SPSS, ArcGis).
Researcher Tools	
PSE	Problem solving environments used by researchers (e.g., R, Stata, Matlab)
Software packaging	Packaging tools used for the distribution of software (e.g., pip, CRAN, conda, RPMs)
Environment packaging	Packaging tools used for the distribution of the computational environment (e.g., QEMU, Docker, Singularity, ReproZip, SciUnit)
Source control	Source control management systems used by some researchers for collaboration (e.g., Github, GitLabs, Bitbuck)

Table 5.4: Reproducibility infrastructure types and examples

Communication infrastructure

With the emergence of online publishing, the journal website has increasingly replaced the journal itself for the communication of guidelines and policies to authors. Where it was once common for new policies to be announced via editorials and included in published issues, today's journal editor typically announces policy changes via blog or other social media (e.g., Twitter) posts and the policies and guidelines are posted only to an unfortunately ephemeral website[2]. Journal websites are used to convey author and reviewer guidelines and present links to submission systems and online journal issues.

Publishing and editorial infrastructure

Publishers and associations maintain their own digital libraries or search engines to provide researchers with access to the online version of published articles. These systems typically implement various features related to accessing supplemental or externally published information and filtering articles based on attributes (i.e., metadata). For data sharing and reproducibility initiatives, digital library platforms are essential for making the link between published papers and externally published materials. Through various means, these platforms also allow readers to determine or discover whether a particular paper has undergone reproducibility review. For example, the ACM Digital Library allows users to filter articles based on their new badging system. The utility of this feature may be unclear for the average user, but is helpful for analyses of policy effect, making it easy to distinguish articles that have undergone different types of review.

Editors and conference organizers rely on editorial and conference management software for the submission and peer review process. The specific software used if often determined by practices within an association or by the publisher. For journals, two commercial platforms dominate, ScholarOne and Editorial Manager. These systems are used the submissions of

[2] This practice presents challenges the type of analysis presented here, as researchers can only turn to the Internet Archive to understand historical changes to policy and editorial procedure.

manuscripts and supplemental information, peer review, and production publication. These reproducibility initiatives are implementing fundamental changes to the peer review that are not yet part of editorial or conference infrastructure. In fact, in all of these cases, communities have had to devise workarounds or implement new processes and tools to support review. This is in part because the reproducibility review is not technically peer review and doesn't follow the same constraints.

Data/code infrastructure

The studied initiatives rely largely on existing data infrastructure. This includes online code sharing and collaboration tools, such as Github (*JASA-ACS*, *IS*, *SC*) and archival research data repositories such as Dataverse (*AJPS*, *JASA-ACS*), OpenICSPR (AEA), Mendeley Data (*IS*), Zenodo (Biostatistics, *SC*, *IS*), ACM Digital Library (*SC*, *TOMS*) and Figshare (*JASA-ACS*, *Biostatistics*, *SC*, IS). In some cases, the selected research data repository is determined by the association (AEA and ACM including *SC* and *TOMS*), publisher (*JASA-ACS*, *IS*) or the journal (*AJPS*, *Biostatistics*). In turn these repositories are more or less integrated with editorial platforms and digital libraries.

Reproduction and verification infrastructure

Existing publishing, editorial, and data infrastructure provides limited support for the reproduction and verification process. The three initiatives with mandatory review processes (AEA, *AJPS*, *SC*) have each implemented custom workflow management and tracking systems for reproducibility review and verification. The AEA uses a custom workflow implemented using the Atlassian JIRA system while the Odum Institute implemented a custom relational database to track curation and verification reports and related information across author resubmissions. These systems are necessary because the curators, reviewers, and verifiers are not part of the traditional peer review process supported in traditional systems such as ScholarOne and Editorial Manager. Both the AEA and Odum also maintain intermediate storage infrastructure to manage versions of submissions during the verification process[3].

For initiatives that support or require code re-execution, computational infrastructure in terms of both hardware and software licenses prove challenging. The AEA and *AJPS* initiatives rely on resources provided by collaborating institutions. The AEA reproducibility initiative is led out of Cornell University and the Labor Dynamics Institute and relies on expertise and resources provided by the University as well as the Cornell Institute for Social and Economic Research (CISER). The *AJPS* initiative relies on computational resources and licensed software provided by the Odum Institute or the University of North Carolina at Chapel Hill. The other initiatives rely on resources provided by the editor or reviewer's host institution. Computational resources vary from laptop and desktop systems to virtual machines or even batch compute infrastructure.

Researcher tools

While not specifically required by any initiatives, there are many tools used by researchers that can simplify the reproducibility or verification process. This is explored further in Chapter 7.

Arguably one of the most important developments in the reproducibility of computational research is the widespread adoption of common scientific problem solving environments (PSEs). PSEs typically provide interactive analysis and visualization interfaces along with high level programming languages and extensible libraries of tested computational methods. Common PSEs include MATLAB, R, Stata, SAS, and SPSS. PSEs provide a common platform and method of packaging and distributing research software that simplifies the verification process. PSEs often provide their own mechanisms for the packaging and distribution of software (e.g., CRAN). Of course not all researchers rely on PSEs and general

[3] Odum is currently developing a new system with funding from the Sloan Foundation to better support the verification process.

purpose programming language and operating systems also provide mechanisms for software packaging and distribution (e.g., pip, Maven, Conda, RPMs).

In some cases, computational reproducibility may require additional information about the versions of installed software or the operating system. Virtualization techniques have made it possible to share entire binary images of the computational environment to improve potential reproducibility in the absence of comprehensive version information. In addition to commonly used methods such as virtual machine or container images (Docker, Singularity, etc), specialized software that also track computational provenance information has been developed (ReproZip, SciUnit, CDE).

5.4.5 Challenges

Interview participants reported a number of common challenges. These include issues of awareness and policy communication; policy continuity in the face of leadership changes; increased burden on authors and editorial staff; computational scale and complexity; gaps in infrastructure; handling non-reproducibility; the use of students; and handling private and protected resources"

Communication and awareness

A common challenge reported by participants was communication of policies and changes to both authors and other initiative stakeholders, including reviewers and editorial staff. One conference chair considered this one of the biggest challenges of their initiative:

> The biggest one by far is communicating what you're actually trying to do...[W]e tried a campaign of education with blog posts and trying to circulate the word through the channels we had, and probably it didn't go very far or something, but just people didn't seem to understand what they were working with. So communicating and describing the process is I think the hardest part, especially across languages. This multilingual environment, the word "reproducibility" has so many different meanings across to non-native English speakers. [4-7]

An editor similarly discussed the challenges of communicating the new review process to associate editors:

> I think the biggest thing that we have to deal with, respect to associate editors is communicating with them how the process is supposed to work and getting them to assign things to us at the right time and to think about us as being part of the process. So that's just the information dissemination task, essentially, to make sure that everybody's on the same page. [3-2]

A lead editor who took over an existing policy and process noted that "I had to learn and be told the most basic features of how it got done." [1-8]

Burden

Another common theme among participants is the increased burden on authors, editorial offices, and reviewers imposed by the new policies and processes. For authors and reviewers, packaging and reviewing the additional materials comes at a cost. For journal and conference leadership, they must take on ownership of the complete process.

On the topic of author burden, several participants noted that it is greater for authors with poorer practices. A managing editor observed:

> [I]t added more work on their end. They had to not only produce the manuscript, but they also had to show their replication code and then they had to make their data available. They had to cross the 't's and dot the 'i's, and it added more time and probably stress for a lot of authors, especially those who didn't have to have a standard of replication and reproducibility. [1-1]

Reflecting on the low-participation rates in a voluntary initiative, one editor noted "I guess maybe that was predictable. Not many people would voluntarily submit to this just for the hassle alone, I think." [4-1]

A lead editor who took over an existing policy and process observed:

> It imposes substantial costs on authors and on editorial offices...I didn't realize how extensive, how detailed, how time intensive the whole process was. [1-8]

When asked to reflect on the challenges of another journal adopting a similar policy, the editor continued:

> You have to manage association or ownership, publishing house ownership issues and stakeholders, which will take some time, you have to figure out what's the best policy, you have to communicate it and given how little most editors are paid and the resources available to them, that is a heavy lift. I mean that's a lot, and I think because the day to day responsibilities for any journal where, now it's different with associate editorial teams, but if you're a lead editor or let's say, just say, if you're the editor of a journal and you're doing more than 500 papers a year, that's a daily task. [1-8]

The overall burden of these initiatives is quite high for all stakeholders involved. Authors must produce more materials, paying greater attention to details that are often ignored, while possibly gaining new skills. Conferences and journals must expand reviewer bases to process this new information, balancing costs and incentives. Editors must define and sustain the overall policy and process – one that is sometimes inherited from a previous administration.

Leadership changes

Planned or unplanned leadership changes are inevitable, as most journals or conferences have rotating editor and chair positions. In this study one initiative (*AJPS*) experienced unplanned editorial leadership changes and four (*Biostatistics*, *JASA-ACS*, *TOMS*, *SC*) had planned changes during the early stages of policy execution. Leadership changes pose a distinct challenge to these initiatives, as there is generally no requirement that policies are maintained.

As noted by one chair: "I think [initiative] is not set up to do a ten year project, which this might be[...]" [4-6]

Computational scale

An additional challenge for most initiatives is the scale and complexity of the computational work undertaken by researchers. For the two initiatives with mandatory reproductions (AEA, *AJPS*), the computational requirements to reproduce research are generally low. However, interviewees from both initiatives report cases of larger scale analysis and how these have been handled. One editor gave the following example:

> 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. [3-1]

Other initiatives receive a mix of submissions. One editor noted that they did not require reproduction because of this concern:

> [T]he challenge that we felt was that, a large fraction of the papers that we get to [journal] use fairly computationally intensive methods. This is going to run for the simulation runs 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. [3-2]

Computational scale – along with issues of reproducibility in parallel and distributed systems – are primary reasons that the *SC* initiative cannot pursue actual reproductions of the results of submitted papers. As a result, the *SC* initiative is less focused on exact reproductions and more focused on building confidence in presented results. The optional Artifact Evaluation (AE) appendix is intended to supplement the AD appendix, encouraging authors to report "meta-computations" to further improve trustworthiness of results.

The two initiatives with mandatory reproduction requirements (AEA, *AJPS*) rely entirely on computational infrastructure provided by host institutions (Cornell, UNC). Other initiatives rely on computational resources provided by reviewer institutions, which has proven problematic. Access to computational resources is a key infrastructure gap in the reproducibility review process.

Infrastructure gaps

A common challenge reported by initiative stakeholders is gaps in infrastructure required to support the expanded peer review process. In three cases (*AJPS*, AEA, *SC*), stakeholders implemented custom systems to address limitations in the primary editorial and publishing workflow. Key limitations reported by interviewees include:

1. Editorial and review software doesn't support reproducibility review workflows and custom workflows do not integrate well with conventional editorial management systems
2. Publisher websites and digital libraries lack facilities to support badging and searching for verified papers or associated materials
3. Reproduction activities require access to computational environments (clusters, large memory machines, specific software versions) including licenses
4. Infrastructure for supporting review of protected or proprietary information

Non-reproducibility

All of the initiatives in this study assess reproducibility post-acceptance or post-publication. These initiatives are largely seen as supportive, where reviewers or verifiers work with authors to meet policy requirements in semi-blind or open review. In several cases, instances of irreproducibility may delay publication, but have not resulted in retraction. *AJPS* has reported that fewer than 10% of submissions pass the verification process the first time. How many of these issues are serious or minor is currently unknown. As noted by an editor involved in another initiative "I'm still waiting for the proposal, not surprisingly, for the manuscript that comes through, that is perfectly reproducible on first pass." [3-1] Aside from publication delay, only the *TOMS* initiative has a stated policy regarding irreproducibility:

> RCR 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.

In pre-publication reviews, the outcome is that the publication of the paper is delayed until issues are corrected. However, this poses additional challenges in post-publication reproducibility assessments. If a paper cannot be reproduced post-publication, does it result in a retraction? One editor noted:

> One I think for sure, before even starting this, I would definitely think more closely about what to do if the research is not reproducible, because I don't think it is fair to just ignore that these things happen. I think this is really, really important. If we really want to make a difference in terms of reproducibility, I think these cases also should be taken into account. [4-2]

Conditional acceptance policies may provide sufficient safeguards to address irreproducibility pre-publication. Of course, subsequent analysis or replication attempts may identify other problems, resulting in adjudication or retraction. However, in post-publication assessments, the reproduction process itself risks the journal highlighting the irreproducibility of an already published work.

Use of students

The full reproduction and verification of computational results presented in a paper is a labor-intensive process through which little is gained in terms of new knowledge. The two initiatives with mandatory full reproduction processes (AEA, *AJPS*) rely heavily on students (undergraduate and graduate) and post-doctoral researchers to conduct the actual verification. The SCC RC is targeted specifically at students reproducing key results of a previous paper. Other initiatives reported considering the use of students as part of the current initiative (*IS*, *JASA-ACS*) but have not done so.

In one case, a journal initially considered relying on traditional associate editors to enlist graduate students for the reproducibility assessment. "We were already asking associate editors to kind of adapt their workflow to do this and to have them then also have to chase down reproducibility reviewers...[i]t was a lot to add to the process already." [3-2] Another journal editor reported considering a change to recruit early-career faculty or postdocs. "Maybe younger people, younger professors, or even postdocs, they're more susceptible to actually get involved in such initiative because they have...more time to work on this." [4-2].

For those initiatives that rely on students, the benefits are clear. The students are typically compensated monetarily and gain additional experience or exposure to state-of-the-art methods. However, there are additional concerns about relying on students to review or verify published research. Interviewees expressed three broad concerns about the use of students. First, community members often do not believe that they are qualified to conduct the assessment of research presented in top-tier outlets. Second, interviewees expressed concerns about bias toward or against individual student reviewers. If their identity was known, it could put job prospects at risk or they might be biased towards helping potential colleagues. Third, there is an ethical issue if they are able to gain an original publication by finding an error in an accepted paper.

Protected and private resources

Most of the initiatives presented here must address challenges of reproducing work that relies on either protected or private information. Central examples include research related to human subjects or commercial interests and may include data, software, or hardware. The question becomes how to assess or verify the reproducibility of published research that relies on protected or private resources. Solutions include the adoption of policies that researchers must provide detailed instructions (access protocols) describing how access these resources.

In two cases, reviewers signed non-disclosure agreements to gain access to proprietary data and computational resources. The *TOMS* RCR initiative allows for guest access to remote systems for the reproduction process. In one case, external reviewers were recruited from the author's institution to conduct the assessment.

5.5 Discussion

To improve the reproducibility of published research, the NASEM committee made the following recommendation:

> 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.

Each of the seven initiatives in this study have implemented either a pre- or post-publication reproducibility assessment process that conforms with the above NASEM recommendation[4]. A central goal of this chapter has been to explain why and how each of the initiatives has defined and implemented its assessment process and the factors that impact their operationalization. These initiatives vary in many ways including the scope and scale of the audit process, who performs the review, what they are reviewing, as well as the resources required or available to reviewers to complete the task. These essential details underlie the "technical and practical challenges" referred to in the recommendation above. The results presented in this chapter highlight several dimensions that contribute to the success of the audit process

[4] This is not to imply that the NASEM recommendations inspired these initiatives. In facts, the studied initiatives all pre-date the NASEM committee and even influenced the report.

including social (who conducts reviews with what expertise), community (norms, motivations, and historical antecedents), information (policy and artifacts), and technical (tools and infrastructure used by authors, reviewers, and publishers).

Reproducing Results or Reusable Software?

For the seven initiatives, there are two broad motivations. First, there are those communities that are concerned primarily with the quality of reported results caused by general carelessness or the potential misapplication of methods. Second, there are those communities that are concerned with the quality of the software produced by researchers. In the first case, initiatives are focused on transparency in the interest of enabling additional scrutiny, likely through replication. In the second case, initiatives focus on the quality and trustworthiness of the software primarily for future re-users or application to new problems. The assessment of one may not benefit from the assessment of the other. These two broad motivations also shape what is required of authors and expected of reviewers. The first results in policies that authors must provide the code and data used to produce results in the paper, regardless of how they are packaged or organized. The second focuses on packaging and distribution of reusable code and data and is generally less concerned with the strict reproduction of results. For example, the AEA and *AJPS* initiatives are more focused on results reproducibility while the *Biostatistics*, *JASA-ACS*, and *TOMS* initiatives are more concerned with the creation and transfer of reusable and trustworthy software. Communities interested in adopting assessment processes need to be clear about the differences in reviews focused on results reproducibility and reviews based on code or data reusability. JOSS[5] provides an example of criteria that are focused more on the quality and reusability of the packaged software than the verification of scientific results and claims.

[5] `https://joss.readthedocs.io/en/latest/review_criteria.html`

(Computational) Reproducibility *of What*?

Recall the NASEM definitions of *computational reproducibility* as “obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis” and *transparency* as “the extent to which researchers provide sufficient information to enable others to reproduce the results.” Radder [225] helpfully differentiates between *reproduc**tion*** and *reproduc**bility*** in experimental research where *reproduction* refers to actual events in the past or present of reproducing an experiment and *reproducibility* is the (fallible) possibility of reproducing the experiment. Each of these initiatives directly operationalizes these concepts within the narrow context of computation. On the one extreme, the *IS* invited reproducibility paper is a full computational reproduction with an extension to test the “workability,” extensibility, and robustness of the provided code. The *AJPS* verification policy ensures computational transparency through a full computational reproduction and confirmation that the results match those presented in the paper. On the other extreme, the *JASA-ACS*[6] and *SC* AD/AE initiatives determine the reproducibility of results in submitted manuscripts through an assessment of the completeness of an associated appendix. While reviewers may execute the provided code, they are not required to do by policy.

As first observed by Dewald, Thursby, and Anderson [74] and reconfirmed by both the earlier SIGMOD repeatability experiment [174, 173, 34] and current *AJPS* initiative [135], materials provided by authors are typically incomplete and inadvertent errors are commonplace. These examples suggest that assessments without actual reproductions will continue to result in artifacts that have oversights an errors that potentially impact reproducibility and understandability. It is unclear – and worthy of study – how many of these issues cannot be resolved by a future researcher and that might truly impact their ability to use the existing artifacts in the service or reproduction or replication.

[6] The *JASA-ACS* reviewer guidelines explicitly state reviewers are not required to run the code. Although they may choose to do so, there is no indication whether they did on the reviewed artifacts.

The ability of a journal or conference to undertake full or partial reproductions will depend on the scale and complexity of computational components of reported research. In the few cases that implement complete reproductions, the research either generally requires small-scale computational resources (e.g., AEA, *AJPS*) or the initiative determines which papers are reproduced (e.g., *IS*, SCC). For those initiatives where scale and complexity are generally high (*JASA-ACS*, *SC*), full reproductions are not considered possible and alternative modes of assessment are required. Full reproductions also require access to suitable computational resources, which poses additional challenges to journal editors and conference organizers. For large-scale research, the *SC* AE appendix presents one approach where, through narrative, researchers describe additional steps taken to improve the trustworthiness of results. Another alternative is the scientific "reduction test" proposed in [155] to demonstrate that the published code and data are working properly, although not a direct verification of results.

(Computational) Reproducibility *By Whom*?

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

These assessments do not seem to require specific knowledge of the research domain, aside from methods, but do often require a deep technical expertise and attention to detail. Since reviewers or verifiers are not assessing the correctness of the computations or considering the theoretical framework of the research, one questions whether it's necessary to rely on peers in the community to do this work. In their discussion of replications in economics, Mirowski and Sklivas [187] suggest that increasing replications "might just involve going around the entire structure of costs and benefits by requiring apprentice empiricists (perhaps at the graduate student level) to attempt replication of one or more articles in the same way they are now required to do theses." If the goal is to verify the reproducibility of all papers, apprenticing advanced graduate students may present a solution to the incentive problem. They stand to gain the most from the experience and exposure to new research, whether compensated or not. However, the use of students does present potential problems in terms of accountability and bias. Initiatives leveraging students will be faced with responsibility for managing the work that needs to be done and confirming its quality. However, the stakes in this case seem quite low, since they are only confirming or disconfirming their ability to re-execute computations. However, students present a number of challenges concerning sustainability (turnover) as well as sources of bias.

Infrastructure Requirements

Over the past two decades, much progress has been made in establishing infrastructure for sharing research artifacts associated with publications, primarily through research data archives. However, any journal or conference attempting to implement the NASEM recommendation 6-4 will need to recognize significant gaps in infrastructure that have not yet been addressed. Existing editorial management tools used for traditional peer review are not well-suited for the types of review and information required for the assessment of supplemental artifacts. As a result, three of the initiatives studied here (AEA, *AJPS*, *SC*) have undertaken the development of new tools in support of the review process. Additionally, anyone interested in performing actual reproductions will need to assess the availability of suitable computational resources. Both AEA and *AJPS* rely on resources provided by host institutions (Cornell, UNC). While many institutions may provide access to such resources, they may not be easy to access for the purposes of journal audits. Funding agencies, such as NSF, could potentially provide access to required resources through initiatives such as XSEDE [263] or Jetstream [248]. Today, it is repositories that provide the core infrastructure for packaging data and code for sharing (e.g., OpenICPSR, Dataverse, Mendeley Data, Zenodo). Other tools that support more granular packaging research artifacts for reproducibility audits are perhaps useful, if generally adopted by the journal or community, and if they reduce burden on authors and reviewers.

Policies, Guidelines, Workflows, and Checklists

An essential component of any reproducibility initiative is a clearly articulated and defined 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. In the past, policies were communicated through editorials published in the associated journal. More recently, editors have turned to less formal modes of communication, through journal websites and blog posts. Because of this practice, the study reported here relied on the Internet Archive to identify and obtain copies of previous policies. Ideally, any policies, workflows, guidelines and checklists would either be published or archived with a persistent identifier, linked to the publications that were reviewed under them. This way it would be possible to identify the particular policy and workflow that was used to assess a paper.

Badges are seen as a way of incentivizing authors, but also providing a mechanism to systematically identify which papers have undergone reproducibility assessment. In the absence of badges, initiatives should consider other ways to clearly identify which papers have undergone review and under which policy. This will help in any future studies attempting to assess initiative impacts.

Measuring Impact

The original *JMCB* study was conducted as experiment on how changes in journal policy impact the availability and quality of research materials [74]. The reproducibility initiatives studied here are largely grounded in experience and intuition, not evidence. Most of these initiatives consider the practice of computational reproducibility to be beneficial and are more concerned with measuring the negative impacts on publication and review than the positive impact of the initiative. Anecdotally, two journals (*AJPS*, *Biostatistics*) reported increases in impact factor, but these cannot be tied directly to changes in policy. The question remains open as to whether these reproducibility initiatives actually result in improvements to research quality or rigor or have other desirable effects on researcher behavior. Journals and conferences looking to adopt the NASEM recommendation should consider how to measure the potential impact of the initiative on research quality. Perhaps the easiest way is to instrument the process and expose the resulting data for analysis. In this sense, opening the black box of peer review to future researchers. 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 downstream 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.

Trust or Verify

A common theme for all of these initiatives is also part of the title of this dissertation: "trust, but verify." In all seven cases, the assessment process occurs after the paper has been accepted and generally has no direct impact on the main peer review process. The reproducibility assessment may delay publication and, in all but one case (*TOMS*), will not result in rejection. The primary peer review process follows traditional practice and focuses on the content of the manuscript, not the associated computational artifacts. The assessment process and associated reports are secondary to the peer review process and are largely viewed as confirmatory or supportive. In this sense, the initiatives trust that the authors of accepted papers can and will provide the materials necessary to reproduce reported results while the assessment process serves enforce and verify that they have done so, to varying degrees of detail. This is consistent with earlier findings that, under current incentive structures, authors will not voluntarily provide these materials and if they do they are likely to be undetected ambiguities, errors, and oversights [187, 90]. While these issues are seen as largely correctable and rarely impacting core findings, they do affect the understandability and reusability of artifacts by future researchers.

Each of these initiatives represents and expansion of the peer review process that increases the burden on authors, editors and reviewers in the interest of improving quality, rigor, and trustworthiness of results. In a critique of computational reproducibility policies, Drummond [83] 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." Leek and Peng [161] argue that computational reproducibility is insufficient to address problematic research and instead argue for a "prevention" approach through increased education. Resnik and Shamoo [227] argue that reproducibility is an ethical problem. Many of the initiatives studied here have been accompanied by changes to research ethics guidelines by broader academic associations.

In a system that relies on trust and integrity, the question remains whether the effort would be better spent on a campaign of education about the responsibilities of researchers instead of the example of enforcement policies during the publication process. In response to the question of whether journals should be responsible for reproducibility, Jacoby et al [135] conclude that they should. However, we have no specific evidence whether the "prevention" or "medication" approaches truly have the desired effect.

In this chapter, I have explored what led each of the seven initiatives to implement their computational reproducibility policies and the many factors involved in their operationalization. I developed and compared key characteristics of each initiative to better understand choices made in the implementation of policies and workflows. I have answered the central questions of *what* is being reproduced and *by whom*. I will conclude with a few recommendations for communities considering the implementation of similar policies and workflows.

Chapter 6

What makes a research artifact computationally (ir)reproducible?

6.1 Introduction

Reproducibility checklists abound. From the 1986 *JMCB* study to Claerbout and Karrenbach's "reproducible research" [58], from King's "replication standard" [145] to Sandve et al's "10 Simple Rules" [235], over the years many research communities have proposed guidelines for what it means to publish reproducible computational research. In Chapter 5 I looked at how the seven initiatives at the center of this study operationalize the reproducibility assessment process. In this chapter I investigate the question (RQ2) *what are the characteristics of research artifacts that make them computationally reproducible (or irreproducible)?* through the comparison of initiative policies, guidelines, and checklists. Each initiative provides detailed guidance to both authors and reviewers for what they consider to be important to the reproducibility assessment process and future reproducibility of published artifacts.

This chapter is organized as follows. In the next section, I briefly summarize the methods used to address the research question followed by a review of related literature. This is followed by the results of the qualitative coding and analysis process and a discussion of how identified factors relate to the reproducibility assessment process.

6.2 Methods

The research question is addressed through qualitative coding and analysis of documentary artifacts from the seven reproducibility initiatives [236]. The complete set of documents analyzed for this chapter are listed in Appendix D and includes policies, guidelines, checklists as well as submission forms, appendices and editorials written by initiative stakeholders. The codebook developed for this analysis is available in Appendix C. All qualitative analysis was conducted using ATLAS.ti (v8.4.4).

6.3 Related Work

In their 1986 *JMCB* study, Dewald, Thursby, and Anderson conclude that journals should:

> require the submission of programs and data at the time empirical papers are submitted. The description of sources, data transformations, and econometric estimators should be so exact that another researcher could replicate the study and, it goes without saying, obtain the same results. [74]

Computational reproducibility – or replication in their terms – requires the software and data used by the original authors. The versions of software should also be provided as an "audit trail" to allow future researchers to "trace bugs in the programs, changes in algorithms, and related difficulties." Users of large or proprietary datasets or confidential programs, they contend, should provide versions, identification numbers, and access dates. For long term reproducibility, they suggest that journals maintain centralized archives for the distribution of programs and datasets. Complex and large-scale models, such as the MPS model[1], would likely present additional difficulties for future reproduction.

Many of the elements of the *JMCB* recommendations remain relevant to discussions of computational reproducibility today. Reproducibility requires software (including versions),

[1] The MPS model required access to an IBM VM/370 computer with "two computer tapes containing more than 2500 files of programs and data"

data (including identifiers and access dates), special handling for proprietary and confidential artifacts. Reproducibility will be impeded by studies relying on complex and large-scale computational models.

King's *replication standard* was influenced by the *JMCB* study [145]. He defines a *replication data set* as containing "all information necessary to replicate empirical results." This includes the original data, programs, records, extracts of publicly available data (or directions to obtain them), and a "readme" file describing how to reproduce the numerical results. According to King, data could be limited to the subset of variables and observations used to produce the published results. For long term replicability, he recommends that the final datasets should be made available publicly, ideally through professional archives, with reference to the original publication.

McCullough, McGeary and Harrison [179] expand on these general requirements with additional recommendations about standard file formats; distinguishing between primary and analysis datasets; including operating system and version information; and indicating which programs correspond to which results. McCullough and Vinod previously reported extensively on problems with the numerical reliability of statistical and econometric software [177, 178, 181]. To them, software version information is key to the identification of results impacted by lower-level software errors. Many of these same recommendations are reaffirmed years later by Chang and Li [47].

Stodden [250] considers reproducible research artifacts within the framework of licensing and copyright. Building on Gentleman and Temple Lang's [105] work, she identifies the elements of a *research compendium* to include the paper (including source), data (including documentation and processing code), experimental workflow (including code and documentation, parameters, operating system dependencies), results (figures, data, and associated documentation), and auxiliary materials (e.g., for presentation on the web). She proposes the *reproducible research standard* (RRS) to encourage sharing for "subsequent use and citation" through copyright and licensing. She recommends that, by default, authors release the compendium to the public domain using the BSD license for code, CC-By for media components, and Science Commons Public Domain Designation for data.

Alvarez, Key, and Nunez [2] reflect on their experience enforcing a replication policy for the journal *Political Analysis*. In addition to the recommendations above, they propose that authors provide details about the computational workflow, system requirements (e.g., cores), and recommend the creation of software packages (for example, using the CRAN format and distribution network). They make explicit recommendations about relating results to tables and figures in the original manuscript as well as the optional inclusion of intermediate outputs.

In high-performance and parallel computing, Höfler and Belli [127] consider the information required to reproduce computational experiments. Beyond software, operating systems, or compiler flags, they note a number of factors that impact reproducibility in the absence of access to the target system. For example, hardware details; network details (topology, latency and bandwidth); allocation policies; system state (warm/cold cache) all factor into the reproducibility in performance studies.

All of these requirements can be found in the policies of the seven initiatives studied here.

6.4 Results

This section reports the results of the cross-case comparison of the seven reproducibility initiatives with respect to the question of what makes a research artifact reproducible (or irreproducible). Table 6.1 summarizes the high-level code groups developed during the qualitative coding process. Over 70 codes were identified in seven broader code groups through repeated coding (see Appendix C.3). In this section, I detail the results of the coding process through a detailed look at each code group.

Code	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

Table 6.1: High-level qualitative codebook categories developed for coding of policies, guidelines, and checklists

6.4.1 Reproducibility

As discussed in Chapter 5, each of the seven initiatives operationalize the assessment of computational reproducibility in different ways. In some cases, assessment requires the full execution of provided workflows and comparison of results to those reported in the manuscript. In others, reviewers assess only the availability and completeness of materials in support of potential future reproductions without requiring re-execution. In some cases, reproduction requires installing software dependencies on a new target system while others allow for access to the original target system. These differences in operational workflows and constraints in part determine many of the factors each initiative considers important for reproducibility. Descriptions of how each initiative characterizes the concept of reproducibility through policy and guidelines documents are provided in Table 6.2. These are important because how the initiative operationalizes the assessment determines in part the characteristics that they deem important for reproducibility.

Initiative	**Description of reproducibility assessment**
AEA	[W]ithin reasonable limits of time and computing resources, we will run your code, and verify that the results produced by your code and data correspond with the publishable results in your article.
AJPS	[M]aterials will be verified to confirm that they do, in fact, reproduce the analytic results reported in the article.
Biostatistics	An article is designated as reproducible if the AER succeeds in executing the code on the data provided and produces results matching those that the authors claim are reproducible.
IS	The goal of the review process is twofold: (i) verify if the results presented in the paper can be reproduced (i.e.: verify if the claims in the paper can be confirmed), and (ii) see how robust the results are to changes in the experiment configuration (i.e.: verify if the software is usable enough to allow others to benefit from it).
JASA-ACS	[W]ithout having run the code, do you have any concerns that the code would not reproduce the key results? Based on having run the code, did the workflow allow you to reproduce the key results?
SC	There should be enough information provided so that a 3rd party could reasonably be expected to replicate or request access to the same modifications used in the experimental setup.
TOMS	Replicated Computational Results (RCR) review is focused solely on replicating any computational results that are included in a manuscript.

Table 6.2: Summary of reproducibility assessment process by initiative

6.4.2 Documentation

Six of the seven initiatives have specific guidelines related to documenting the computational workflow[2]. For non-experimental research (e.g., hypothesis testing), this includes the list of files and the order in which they need to be run; a master script; or other workflow submission protocol. For experimental work, the documented workflow must also include the complete experimental protocol, which may require steps beyond the computational setup. AEA and *AJPS* require a top-level readme file and manifest describing all of the artifacts provided and their role in the reproduction process. Several initiatives require documentation of parameters, including specific random seed values, used during runtime.

[2] I excluded the *TOMS* RCR initiative since under the section "Independent Replication" it provides no specific guidance or requirements for specifying the computational workflow, only the "sufficient description of the computational platform." [122].

6.4.3 Software

In all seven cases, assessment of computational reproducibility requires access to the code, programs, or other software artifacts used to generate the results reported in the paper. Accessibility does not require open access, particularly as the reproducibility assessment process occurs in most cases pre-publication. Authors may provide access to the software by depositing it an initiative-specific repository (AEA, *AJPS* , *IS*), general-purpose repository (*SC*), as supplemental material during publication (JASA-ACS , *Biostatistics*), or even via direct transfer to reviewers or by providing access to a system with the software installed (*TOMS*). The *SC* initiatives include additional provisions for handling proprietary or closed-source software. With respect to software licensing, only one initiative (JASA-ACS) includes guidelines related to licensing for reuse.

6.4.4 Data

Six of the seven cases provide guidelines related to accessibility of data used to generate the results reported in the paper. The TOMS initiative, which is concerned primarily with software, has no specific guidelines for handling data. As with the software, accessibility does not require open access and four initiatives provide specific guidance for proprietary or confidential data (AEA, *AJPS* , *IS* , *SC*). Three initiatives (AEA, *AJPS*, *JASA-ACS*) provide specific guidelines for data documentation, including codebooks or similar metadata, as well as requirements for accurate variable labeling in data files. Three initiatives (AEA, *AJPS*, *JASA-ACS*) provide specific guidelines for citing source data and versions. Four initiatives (AEA, *Biostatistics*, *IS* , *JASA-ACS*) provide guidelines on data licensing, specifically confirming that authors have rights to redistribute data present in a submission. Two initiatives (AEA, *AJPS*) include guidelines related to access to source data used to prepare datasets used in analysis.

6.4.5 Environment

The notion of the computational environment broadly refers to all relevant system aspects that underlie the primary research code and data. This includes the dependent software (applications, libraries, versions, and settings); operating systems and versions; compiler version and settings; required hardware and versions; required resources (disk space, memory, cores, running time); and details of the runtime environment (single user, hot/cold cache, process pinning). In some fields, aspects of the environment including both software and hardware may be proprietary or limited access.

How information about the environment is conveyed for the reproducibility review process differ. For all initiatives, relevant details about the environment are conveyed primarily through textual descriptions provided in README files, documentation or required appendices. The *SC* initiative permits authors to specify system names and dates used (e.g., OLCF Summit July 2018) when "specifications, configuration, and other relevant details are can be reasonably expected to be publicly available for the next ten years." While *SC* does not perform actual reproductions, for the *TOMS* initiative, access to a target system is an acceptable environment for reproduction. *IS* recommends that authors provide the environment via other packaging mechanisms including virtual machine or container image (e.g. Docker) or using reproducibility-aware tools, such as ReproZip. These preserve, to some degree, the relevant "bits" required for reproduction. The *SC* initiative also provides its "author kit" to enable researchers to capture relevant information about the runtime environment.

6.4.6 Experimental context

Three of the seven initiatives (AEA, *IS* , *SC*) provide specific guidelines related to experimental setup. For strictly computational experiments, as in the case of *IS* , and *SC* , the primary goal of "computational reproducibility" is experimental reproducibility. In these cases, the computational environment often plays a key role in reproducibility, as described above. Computational experiments often rely on benchmark programs and datasets or established metrics used for reporting. For the AEA, the experimental context will likely be external to the computational elements of the research. Therefore, documentation of experiments will require information such as human subject selection and exclusions. These details are not directly related to computational reproducibility, but represent an important distinction between experimentation in computer science versus experimentation that relies on computational analysis.

6.4.7 Results

For all of the studied initiatives, the results that are the target of reproducibility assessment are those presented in the manuscript. The final version of the paper contains results in the form of tables, figures, and in-text analytical claims. Only the *Biostatistics* initiative requires that the results be provided as "a 'target' file (or files) containing the results which are to be reproduced." The AEA, *AJPS* , and *JASA-ACS* initiatives require that the relationship between provided code/data and the figures and tables in the paper be specified (i.e., results provenance).

6.4.8 Publishing

Four of the seven initiatives (AEA, *AJPS*, *JASA-ACS*, *IS*) require that artifacts required for computational reproducibility be made available via archival repositories, except for proprietary and confidential artifacts. Two initiatives (*Biostatistics* , *SC*) encourage authors to make artifacts available via general purpose repositories such as Zenodo, Figshare, or Dryad. *TOMS* has no requirement for publishing RCR artifacts, although Algorithms are published to *CALGO* as software. *TOMS* and *IS* reproducibility initiatives result in publications (RCR report and reproducibility paper) that are published in the journal).

6.5 Discussion

The NASEM recommendation 4-1 (see Appendix A) defines the information that authors must provide to ensure the reproducibility of their computational results. As discussed in Chapter 5, recommendation 6-4 states that journals should consider ways to ensure computational reproducibility through audit processes. The results presented in this chapter suggest that the information required of authors is often determined by how journal operationalize the assessment process. Broadly speaking, the NASEM 4-1 requirements are similar to those of the seven initiatives and consistent with previous recommendations. The broad categories of documentation, software, data, and the computational environment can be found in many recommendations for requirements for computational transparency and reproducibility. Inclusion of the experimental context and results provenance, while less common, are certainly not unexpected. While the details may differ, I conclude that the general factors seen to contribute to computational transparency and reproducibility are similar enough to suggest that a single set of guidelines could be developed to meet the needs of different fields with varying degrees of granularity depending on local requirements. To encourage such a set of guidelines, in the next section I identify what I view as the core factors to computational transparency and reproducibility and discuss some of the key differences in these factors across the seven initiatives.

6.5.1 Core factors

There are four key factors affecting computational reproducibility: documentation of computational workflow, accessibility of precise versions of data and software used, details about the computational environment, and long-term accessibility of resulting research artifacts. The remaining factors more directly impact understandability and re-usability of artifacts which, while important, are not expressly required to conduct a reproduction but contribute to research transparency and trustworthiness.

Without putting effort into forensics, in only the simplest of cases can one reproduce results in the absence of a clear workflow. Documenting or automating the steps taken reduces the potential of error on the part of the reviewer or reproducer. Capturing the workflow in a script or automation framework further ensures that the same workflow used by the author is used for the reproduction.

Using the wrong version of software or data may result in irreproducibility of results. The precise specification of data and software versions used as well as citations or clearly documented access protocols for external sources will increase the likelihood of reproduction. Similarly, different versions of software or hardware dependencies may contribute to irreproducibility of results. In this case is it important to document in sufficient detail the computational environment used in the analysis. This will be discussed in more detail below, but describing or capturing details of the computational environment present several challenges.

Many of the other factors identified in the initiative policies and guidelines are beneficial but less essential for reproducibility assessment. These include data documentation, results provenance, sensitive information checks, license and copyright information and long-term archiving.

For originally collected data, several initiatives require detailed and accurate codebooks describing included variables. While not strictly required for computational reproducibility – the reproduction can occur in the absence of documentation as long as the data are available – such documentation increases the understandability and quality of the provided materials, and does enable replicability.

Checking for sensitive information or author permission to redistribute software or data are perhaps important to the curation of research artifacts but do not contribute directly to reproducibility. Similarly, licenses and copyright contribute to irreproducibility in terms of access to information, but more likely affect the re-usability of provided software, data, and content.

6.5.2 Reproducibility and Protected or Proprietary Resources

If computational reproducibility and assessment of reproducibility require access to the data, software, and computational environment, then restricted-access to these resources presents an impediment. While reproducibility is often discussed in the context of open access, the studied initiatives recognize many legitimate reasons for restricted access to data, software, and even hardware. Concerns can be broken down into three broad categories: 1) protected/confidential information (e.g., human subjects), 2) proprietary resources, and 3) intellectual property.

Protected or confidential data are common in the social and medical sciences. This includes, for example, information protected by the Institutional Review Board (IRB), Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability an Accountability Act (HIPAA), or other legal barries to sharing. Data that falls under these protections present challenges for both data management and access. For reproducibility assessment, the AEA presents a compelling approach. Authors are required to provide detailed access protocols as part of their artifact submissions that include detailed instructions for how an individual with appropriate permissions can gain access to protected data. While it may not always be possible to do so as part of the assessment process, this presents a better solution than prior exemptions for research relying on protected information.

Ownership issues similarly arise from proprietary datasets, software, and hardware. Vilhuber [266] reports on the increased use of non-public data and software in economics research. In this case reproducibility assessment relies on access to commercial datasets and software, which may come at a cost. As with protected information, the AEA requires access protocols for proprietary data and, in some cases, will attempt to acquire it or request access for the purpose of assessment. In this sense, proprietary data is at times easily accessible, but comes with a cost, and may not be redistributable by researchers that rely on it in their work. This also applies to the software used for analysis. Common PSEs such as SAS, STATA, or MATLAB require licenses in order to re-execute author-provided code, and may require special permissions to even redistribute in binary form when considering preserving the computational environment.

Intellectual property concerns cover copyright, patent, and trade secrets, and are common in commercial research. This is a form of proprietary information, but in this case researchers are less to be able to provide access. One solution provided by the *TOMS* and also used in artifact evaluation via the ctuning.org initiative, is for the reviewer to sign a Non-Disclosure Agreement (NDA)[3] or to require reviewers to conduct reproducibility assessment via systems controlled by the owning organization. One editor discussed the case of using SSH to access a remote resource for review where the reviewer agreed to non-disclosure by accessing the system. Like the access protocol, this provides a mechanism for reproducibility assessment in the case of proprietary data, software, or hardware.

When considering artifact availability, concerns about protected and proprietary resources do not require and all-or-nothing solution. Researchers can always make available all non-protected elements of their research or, as suggested by the AEA approach, provide detailed protocols on how to gain access for someone with the correct permissions. Research relying on protected or proprietary resources must conform to the same standards as research that does not, and only access to those resources is impeded. Stodden [250] proposes the reproducible research standard as a framework for encouraging sharing of reproducible research artifacts through copyright and licensing. This standard can be applied, allowing researchers to distribute those elements of their work that are not protected. For reproducibility, it is most important that the process or protocol for obtaining access is clearly documented, even if few will ever gain access. Organizations should do their best to provide access for reproducibility review. However, we must still recognized that restricted access resources may still hinder future reproduction attempts.

[3] This is also used in the case of student verifiers in the *AJPS* initiative

6.5.3 The Computational Environment: A Matter of Degree

Broadly speaking, the requirement to capture information about the computational environment is considered an important factor in computational reproducibility across the studied initiatives. However, what constitutes the environment and the information that needs to be provided differs widely. As discussed in the previous section, for several initiatives documentation of the version of a specific PSE (e.g., RStudio, MATLAB); operating system and version; and dependent software libraries and versions are generally accepted as a specification of the environment. For others, information about compiler versions and settings; hardware and versions; and even runtime conditions (e.g., single user, hot/cold cache) are equally important.

Hinsen [126] proposes the concept of *software collapse* in scientific computing to describe how software becomes unusable because of changes in a layered set of dependencies. Figure 6.1 illustrates his scientific software stack and associated layers. While Hinsen is concerned with collapse (or rot), his framework is also helpful in understanding the effect of the computational environment on reproducibility. Software higher up on the stack is less likely to be maintained over time and also impacted by changes lower in the stack. Reproducibility initiatives are in general concerned with capturing layer-4 research artifacts along with information about the lower layer dependencies. However, layer-4 artifacts are the least likely to be ported over time as lower layer dependencies change.

Capturing the environment is therefore a difficult problem. Over time, older versions of hardware and software are no longer supported or maintained. Consider the example of the popular open-source Ubuntu operating system. Ubuntu developers releases a new interim version every six months and a new Long Term Support (LTS) version every two years. The interim releases receive nine months of support while the LTS versions receive five years of support, after which they reach end-of-life. End-of-life means no further security, package, or maintenance updates. It may no longer be possible to install dependent software. Similarly, newer operating system releases may no longer support installation of older packages, particularly those that rely on specific versions of layer-1 software.

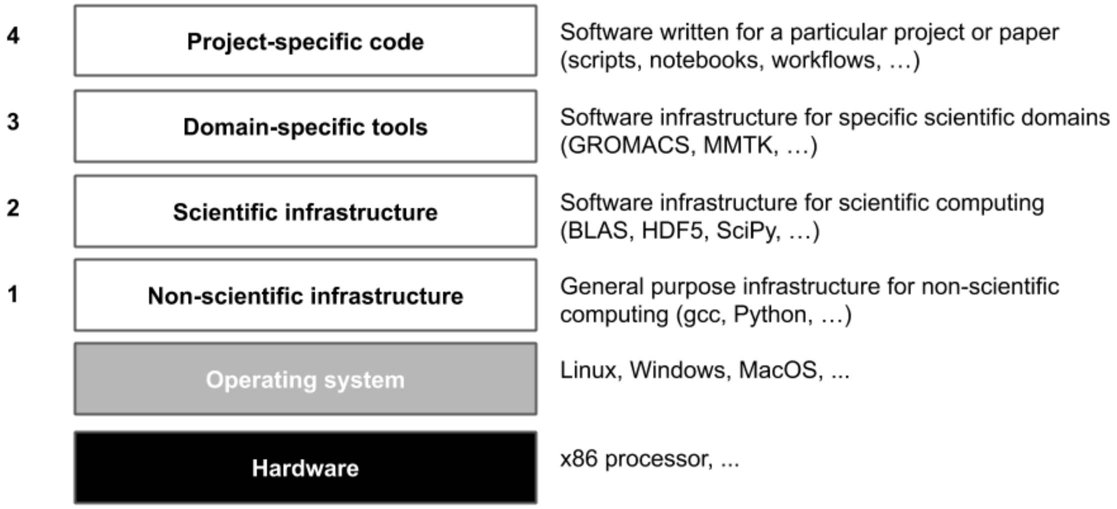

4
Project-specific code
Software written for a particular project or paper (scripts, notebooks, workflows, …)
3
Domain-specific tools
Software infrastructure for specific scientific domains (GROMACS, MMTK, …)
2
Scientific infrastructure
Software infrastructure for scientific computing (BLAS, HDF5, SciPy, …)
1
Non-scientific infrastructure
General purpose infrastructure for non-scientific computing (gcc, Python, …)
Operating system
Linux, Windows, MacOS, ...
Hardware
x86 processor, ...

Figure 6.1: Hinsen's scientific software stack

Computational research happens at all layers of Hinsen's stack. Particularly in computer science, where lower level software and hardware are the experimental environment or the object of study. For research conducted with software at the upper layers, virtualization techniques may provide a mechanism to capture the environment to enable reproducibility for longer periods of time. This way the individual researcher does not need to consider all of the possible dependencies that impact the reproducibility of results. This becomes more complicated when results rely on hardware or underlying runtime state. Other techniques may be required to capture runtime information about the system. However, we can envision multiple fields that would benefit from standardized descriptions at each level. While the social sciences may not need the ability to specify hardware information or runtime states, any subfield of computer science concerned with performance likely would. A comprehensive set of guidelines concerning the computational environment for reproducibility could be used to compose individual journal or conference guidelines *a la carte*. As authors struggle to understand the information required to capture the environment, tools may provide a convenient way to automate this process.

6.5.4 Provenance as a Matter of Trust

In museum curation, the notion of provenance is central to determining the authenticity of an artifact. In computational research, provenance may be the key to establishing trustworthiness of results in the absence of complete re-execution. As will be discussed in the next chapter, automated provenance capture techniques could be used to encapsulate the exact versions of software and data along with detailed information about the environment used in the generation of computational results. This could provide a mechanism to "authenticate" results and establish further trust in the provided materials.

In this chapter I've explored the question of what it means for a research artifact to be computationally reproducible. Based on the seven cases, I identified four key factors including documentation of computational workflow, accessibility of precise versions of data and software used, details about the computational environment, and long-term accessibility of computational research artifacts. I considered other factors that more directly impact understandability and re-usability of artifacts. Finally, I discussed the challenges of documenting or capturing the computational environment and the potential role of automated provenance capture information in increasing trust in reported results.

Chapter 7

Packaging research artifacts for computational transparency and reproducibility

7.1 Introduction

In this chapter, I look at available tools and packaging formats and how they might help meet reproducibility policy requirements. While the initiatives examined in this study all provide guidance on what information authors must include with publication, few require or recommend the use of specific tools, technologies, or packaging formats beyond research data archive platforms.

Over the past two decades, a number of tools have been developed or re-purposed to aid in the creation and publication of computationally reproducible research artifacts. These include tools for literate programming, scientific workflows, provenance capture, record-and-replay, virtualization, as well as general-purpose reproducible research platforms. Work has also been done on frameworks for the organization, packaging, and description of resulting artifacts.

In this chapter I explore the question (RQ3) *what are the key characteristics of tools and packaging formats that enable computational transparency and reproducibility?* I consider how these characteristics relate to the requirements of the seven initiatives and may be used to aid in the reproducibility assessment process.

7.2 Methods

As detailed in Chapter 3, this research question is addressed through qualitative content analysis [236] of a sample of up to 5 verified artifacts from each initiative[1], combined with the analysis of existing tools and packaging formats for the creation and publication of reproducible research artifacts. Combining these sources of information allows me to explore capabilities available to authors and initiatives as well as those actually used today in review and verification processes. The results of this analysis are used for the development of a taxonomy of characteristics of packaging formats used for the representation of computationally transparent and reproducible research artifacts and an abstract model of the research compendium concept. The complete list of artifacts used in this analysis are provided in Appendix F.

7.3 Related Work

Over the past two decades, a number of tools, formats, and conventions have been proposed to support the packaging, distribution, and re-execution of reproducible computational research artifacts. This includes tools for the creation and publication of reproducible documents [23, 163, 232, 238, 278]; research compendia [105, 176]; research objects [24, 26, 242, 51]; tools for virtualization and containerization [31, 56, 158, 113, 206, 262]; scientific workflow and automation systems [71, 96, 101, 273]; provenance capture tools [56, 96, 218]; record and replay tools [50, 190]; and more general conventions or frameworks [140, 175, 223]. Many of these rely on more established infrastructure for version control, research data management, and software distribution. Perhaps the most common and established tools across the seven initiatives for the packaging and distribution of computational research artifacts today is the research data repository. Platforms such as Dataverse, Dryad,

[1] The *IS*, *Biostatistics*, and *TOMS* initiatives all have five or fewer verified artifacts at the time of analysis.

OpenICPSR, Zenodo and FigShare feature prominently in initiative policies and guidelines. In this section, I review examples of the broad classes of tools and packaging formats that are part of the reproducibility infrastructure ecosystem, summarized in Table 7.1.

Packaging tool/format	**Description**	**Examples**
Software repositories	Tools that enable the packaging and distribution of software through common OS or language package repositories	PyPI, CRAN, Conda, Maven, RPM, ...
Data repositories	Tools that enable the publication and archiving of research data and related artifacts and assignment of persistent identifiers	Dataverse, Dryad, Zenodo, OpenICSPR, FigShare, ...
Version control repositories	Tools that enable distributed collaboration, version management, release, and automation tasks.	GitHub, GitLab, BitBucket, ...
Virtualization and containers	Tools that enable packaging and distribution of virtual application images	VMs, Docker, Singularity, CDE, ReproZip, Sciunit ...
Image registries	Repositories for the distribution of virtual machine or container images	GCE, AWS, OpenStack, DockerHub, Quay, Singularity Hub, ...
Problem solving environments	Interactive analytical environments used by researchers to implement their primary research code and workflows. PSEs generally integrate with software repositories, version control repositories, and reproducible document tools.	R (RStudio), Python (Jupyter), MATLAB, STATA, SPSS, ...
Reproducible documents	Tools that enable the creation of dynamic documents that combined narrative, data, and code	ReDoc, knitr, Sweave, Jupyter, RMarkdown, ...
Scientific workflow systems	Tools that enable creation and execution of computational workflows	VisTrails, Pegasus, Kepler, Galaxy, OCCAM, CK, ...
Computational provenance capture	Tools that enable declaration or capture of computational provenance	YesWorkflow, noWorkflow, RDataTracker, ReproZip, ...
Metadata standards	Structured metadata standards for the representation of composite research artifacts	Research objects, RO-Bundle, RO-Crate, ...
Packaging conventions	Frameworks that define specific conventions for reproducible research packaging	Popper, research compendia, rrtools, TIER, ...
Reproducibility platforms	General purpose platforms for the creation and re-execution of reproducible research artifacts, often combining multiple technologies (e.g., virtualization, provenance capture, re-execution, interactive environments, standards, conventions)	CodeOcean, Whole Tale, Gigantum, Binder, ...

Table 7.1: Some reproducible research tools and the infrastructure ecosystem

Software repositories

Operating system, language, and discipline-specific software repositories provide detailed specifications and processes for the packaging and distribution of software. Centralized software repositories and distribution networks also generally maintain archival copies of published packages. Examples include the RPM (RedHat Package Manager) and Yum package managers and associated repositories for Linux; the Conda cross-platform package manager; PyPI for Python; Maven for Java; CRAN for R as well as domain-specific systems such as CALGO (mathematics) and StatLib (statistics). Software packages distributed via traditional software repositories are generally not considered to be part of the scholarly record, although research software is increasingly published to research repositories, such as Zenodo.

Data repository platforms

Research data repository platforms[2] emerged in the early 2000s as researchers, communities, publishers, and funding agencies sought to improve the transparency and reusability of scientific data [112, 207]. Early repository platforms were developed based on technology for publishing grey literature. In addition to being committed to the long-term archiving and preservation of deposited materials and issuing persistent identifiers (e.g. DOI), the repository platforms implement metadata standards for the description of research data (e.g., DataCite, DDI). Repositories are often both installable platforms and production services. Repositories operators offer other specialized services, such as curation and support for sensitive or restricted use information (e.g., OpenICPSR). Repository platforms often also support usage tracking and reporting as well as data citation worfklows. These platforms are increasingly used for publishing and citing not only data, but also software and reproducible research artifacts [224].

[2] Scientific communities have maintained data archives since the 1940s, but software-based platforms for managing research data did not emerge until much later.

Version control repositories

Hosted version control repositories, such as Github, GitLab or BitBucket, provide mechanisms for collaboration, version management, releasing, and automating software-related tasks (e.g., continuous integration). While version control platforms are non-archival, several initiatives recommend that authors use them for the distribution of reusable artifacts. These systems provide capabilities such as cloning or forking that simplify the reuse and extension process. For example, the *JASA-ACS* initiative maintains it's own GitHub organization and creates per-paper repositories. In this sense, GitHub is treated as an archival system for the journal, but also allows for easy reuse via practices common in the community (e.g., forking). The *IS* initiative suggests that authors host their code on GitHub. *SC* provides Git or SVN repositories as an example for packaging artifacts for review, with the caveat that they are non-archival and will not be eligible for certain ACM badges.

Virtualization and containers

Virtualization technology underlies much of modern cloud computing infrastructure. Through abstraction and isolation, the same underlying computational hardware can be used to support multiple distinct virtual machines (VMs), each with separate operating system and installed software. The subsystem that supports virtualization is called a hypervisor, and may be implemented via software or hardware. VM images can be exported in a number of formats and are increasingly viewed as a useful method of distributing and preserving research software [231].

Containerization has emerged as an alternative and more efficient approach over hypervisor-based systems. Containers are generally more resource-efficient, provide a clear and lightweight abstraction for packaging over the full virtual system [243], and are also seen as a possible solution for the preservation of research software [185]. Common container platforms include Singularity [158] and Docker[3]. As with VM images, container images are viewed as a method of distributing and preserving research software, although not designed with this purpose in mind. Container images use a layering technique to achieve storage efficiency, where multiple images may reference the same layers, which are each stored only once. Leveraging this storage efficiency requires the use of specialized software called image registries.

Other approaches to virtualization have emerged from the computational reproducibility community. The Code, Data and Environment (CDE) system was developed to simplify the process of sharing scientific code along with required software dependencies [113, 114]. CDE instruments programs during execution to identify code, data, and environment variables used and creates a self-contained package (or container) that can be easily re-executed. CDE is the basis for the Provenance-to-Use (PTU) [218] and Sciunit [262] systems described below, as well as ReproZip which creates a virtual application [56].

Image registries

Cloud computing platforms such as GCE, AWS, and OpenStack provide ways to store and distribute both VM and container images. Public container image registries such as DockerHub, Quay.io, and Singularity Hub also provide centralized storage and distribution of images, but are considered to be non-archival. Researchers commonly publish complete images to research data repositories, but these are often very large, may exceed file system limits, and do not benefit from the storage efficiency provided by the layering approach.

Problem Solving Environments (PSEs)

As discussed in Chapter 2, the widespread adoption of common analysis environments and languages, such as R, Python, MATLAB, and STATA, have provided researchers with common frameworks for the distribution of scientific scripts and software. Sometimes called

[3] https://www.docker.com

problem solving environments (PSEs) [230], these tools provide a common foundation and mechanisms for extension and the distribution of custom packages that can easily be used by others. None of the initiatives recommend the use of a particular PSE environment, although they are widely used within the different communities. For example, the vast majority of submissions to *AJPS* are written using R or STATA. AEA submissions are written primarily in STATA and MATLAB with some R. *Biostatistics* and *JASA-ACS* submissions are written primarily in R. Still, many researchers rely on other general-purpose programming languages and environments. While not PSEs per se, research leveraging standard languages, compilers, and operating systems (e.g., C/C++, GCC, Linux) offer many of the same benefits.

Reproducible Documents

The concept of "reproducible research" introduced in 1992 by Claerbout and Karrenbach [58] is concerned primarily with the reproduction of the manuscript. Their ReDoc system [238] combined manuscript production with software development best practices to enable rebuilding the paper from source files including document text, code, and data. A key innovation is that the computations required to regenerate the data behind figures and tables could be reproduced.

The notion of literate programming [150] applied in this context led to the development of several systems to better integrate documentation with scientific programming [232, 163, 23, 278, 148]. These systems support the automatic reproduction of documents based on code and data or provide integrated interactive environments that combine narrative and scripting. However, they generally provide no direct facility for packaging and redistribution. The *Biostatistics* initiative recommends that authors use literate programming tools, spefically listing NoWeb or Sweave. The *SC* initiative suggest Jupyter notebooks as a way for authors to document their experimental workflows.

Scientific workflow systems

Scientific workflow systems are used to automate often complex computational pipelines, generally providing reusable workflows that are resilient to failure and portable across execution environments [72]. Large scale workflow systems often require researchers to integrate high-level workflow languages and application programming interfaces (API). These systems can also provide detailed provenance information about how results were derived [183]. Prominent among reproducibility-focused scientific workflow systems is VisTrails [69, 97, 96]. VisTrails differs from other scientific workflow systems (e.g., Kepler [167], Galaxy [107], and Pegasus [71]) because of its focus on provenance of both workflow executions and workflow specifications. It enables tracking the evolution of the computational workflow throughout the discovery process as advanced visualization techniques for exploration and comparison. Workflow systems such as Pegasus [71] are more widely used for the automation of complex failure-tolerant computational pipelines, such as the LIGO Collaboration [72]. Collective Knowledge (CK) is a workflow and automation framework developed specifically for reproducibility and collaboration in systems research [100] by leaders in the cTuning community. Only the cTuning community provides specific recommendations for workflow systems.

Computational provenance capture

In spite of the many features of workflow systems, many researchers still rely on custom workflows implemented in common scripting languages (e.g., R, Python, MATLAB [183]. Several tools have been developed to provide some of the features of larger workflow systems. YesWorkflow [183] enables researchers to annotate scripts across multiple languages to “reveal the computational modules and dataflows.” Tools like RDataTracker [164] and noWorkflow [191] can be used to instrument specific languages to provide runtime provenance information, which in turn may be used to derive workflow graphs. Several tools have also tried to combine provenance information with containerization to capture the computational environment.

Provenance-to-Use (PTU) extends CDE to capture system level provenance information [218]. CDE and PTU rely on features of the Linux operating system and are therefore limited to Linux-based research applications. Sciunit [262] builds on both CDE and PTU to create publishable, re-executable 'research objects' that are both provenance-enabled and versioned. ReproZip [56] similarly creates a self-contained package for computational experiments that includes code, data, software dependencies, and system-level provenance information. It was developed by leadership in the *IS* initiative and the recommended packaging format for the SIGMOD reproducibility review. ReproZip is recommended for packaging artifacts for the *IS* initiatave.

Metadata standards

As discussed above, archival repository platforms support a number of metadata standards related to research data description, sharing, and discovery. Most modern repositories support the DataCite standard in some form. Domain-specialized repositories typically support related metadata standards such as DDI in the social sciences or EML in ecological and environmental sciences. Repositories increasingly support standards for the representation of provenance information, such as W3C prov or ProvONE. Taking a linked data approach, Bechhofer et al [25] introduced the *Research Object* (RO) as a way of describing aggregations of resources used in experiments. ROs present a conceptual model and standard that can be used to describe complex aggregations of resources across organizational boundaries. The nascent RO model underlies infrastructure tools including myExperiment [70], Big Data Bags [51] and the Whole Tale platform [52]. Initiatives implicitly recommend metadata standards through chosen repository platforms.

Packaging conventions

Gentleman and Temple Lang [105] proposed the concept of the *research compendium* as a container for the artifacts required to reproduce the results and reuse the computational methods presented in a paper. Gentleman was an author of the R language [133] and the compendium concept continues to be adopted within the R community (e.g., [176, 196]) and underlies systems such as ResearchCompendia.org [256]. Marwick et al [176] describe the three principles of research compendia as 1) organizing files according the the prevailing convention in a discipline; 2) clearly separating data, method, and output; and 3) specifying the computational environment. They introduce the `rrtools` package for the creation of compendia in R.

The Popper convention [140] coming out of the systems research community proposes a model for applying software development practices – commonly referred to as "DevOps" – to scientific computing. The Popper framework combines version control, package management, experiment orchestration (aka workflow), infrastructure provisioning, as well as continuous integration and regression testing. In this sense packaging research for computational reproducibility is closely aligned to software development practices.

Binder[4] began as a way to specify the computational environment required to run Jupyter notebooks contained in a GitHub repository. Researchers conform to the Binder convention by providing a well-defined set of configuration files that can be used to dynamically build and run a Docker image. Binder is widely used for education and training environments and increasingly as a method of packaging research for computational reproducibility.

The CodeOcean (CO) platform implements what they call a "capsule" that combines the code, data, and computational environment for computational reproducibility. Each capsule is based on a customizable Docker image from a set of images maintained by CO. In addition to the code, data, and environment, Capsules include basic descriptive metadata (author,

[4] `https://mybinder.org`

title, description), reference to publication and license and copyright information. Capsules can be exported as a zip archive where the image can be "pulled" from the CO platform or built locally.

The Whole Tale (WT) platform defines the "tale" format that also combines the code, data, and environment for computational reproducibility. WT leverages the repo2docker component underlying the Binder system and combines this with an expansion of the Research Object Bundle specification, serialized using BagIt. Tales also include basic descriptive metadata and license information. The project is expanding the system to support inclusion of computational provenance information.

Reproducibility platforms

Several platforms have been developed to supplement traditional research data repository infrastructure to enable re-execution of computationally reproducible artifacts. These often combine many of the tools and capabilities listed above. For example, CodeOcean, a widely adopted commercial platform, provides support for popular computational environments including PSEs and literate programming environments, the "capsule" packaging format, preservation of the environment via Docker images and specifications. The Whole Tale platform provides these same capabilities while also adding support for automated provenance capture.

7.4 Results

This section reports the results of the analysis of initiative policies, guidelines, and checklists along with the sample of 27 verified artifacts with respect to the capabilities of available technologies in support of computational transparency and reproducibility.

7.4.1 Initiative packaging recommendations and requirements

As detailed in Chapter 6, each of the seven initiatives provide substantial guidance on the information that must be provided by authors, but they provide limited guidance on how it should be packaged for dissemination. Table 7.2 summarizes the packaging tools and platforms recommended or required by each initiative. Despite over two decades of the development of tools and infrastructure in support of computational reproducibility, current initiatives rely on few. Recall that the Dataverse repository is specifically designed for the sharing and dissemination of "replication datasets" [147]. In this sense, the repository itself can be viewed as a primary mechanism for packaging and distribution of research artifacts. The ReproZip tool recommended by the journal *IS* was developed for use by the database community by leaders in the IS initiative [56].

Initiative	**Recommended packaging formats/tools**	**Recommended and required repository platforms**
AEA	Not specified	OpenICSPR
AJPS	Not specified	Dataverse
Biostatistics	NoWeb Sweave	Zenodo, Figshare
cTuning	Docker, VM images, zip/tar archives	Zenodo, FigShare, Dryad
IS	VM images, Docker, ReproZip	Mendeley Data
JASA-ACS	Not specified	Dataverse, Dryad, Zenodo
SC	Not specified	Not specified
TOMS	Not specified	Not specified

Table 7.2: Packaging formats and platforms required or recommended by each reproducibility initiative

All but one initiative (*TOMS*) require researchers to deposit artifacts in an archival repository. Only two initiatives provide any guidance on packaging formats or tools. *Biostatistics* encourages the use of literate programming environments, such as Sweave or NoWeb. *IS* recommends packaging the environment via VM images, Docker images, or using ReproZip, a tool developed specifically for the databases community. These packaging formats are discussed in greater detail in the next section.

7.4.2 Verified artifacts

This section presents the results of the analysis of 27 verified artifacts from six of the seven initiatives[5]. The complete list of artifacts reviewed is provided in Appendix F. Table 7.3 summarizes the tools and formats used by artifact authors in the context of the seven initiatives. Only those tools or formats used are presented.

[5] Supplemental materials for the *Biostatistics* initiative are no longer accessible and could not be reviewed.

Initiative	Artifacts	Repository	Version Control	Virtualization	PSE	Literate programming	Provenance capture
AEA	5	OpenICSPR (5)	NA	NA	STATA (4), MATLAB (1), Python (2)	LaTeX (1), iPynb (2)	NA
AJPS	5	Dataverse (5)	NA	NA	R (3), STATA (3)	NA	NA
IS	3	Mendeley Data (3)	Github (3)	Docker (2), Reprozip (2)	NA	NA	ReproZip (2)
JASA-ACS	5	Institution (1)	Github (4)	NA	MATLAB (1), R (4)	RMarkdown (1)	NA
SC	5	Zenodo (4), DOE (1)	GitLab (1), Github (1)	Docker (1)	NA	NA	NA
TOMS	4	OSTI (1), NERSC (1), Zenodo (1)	Github (4)	VirtualEnv (1)	NA	NA	NA

Table 7.3: Packaging formats and tools used by authors

As previously discussed, research data repositories are widely used by authors for dissemination of both data and software. This is due in part to initiative requirements, as AEA, *AJPS*, *IS*, and *SC* all mandate the use of an archival repository[6]. While the *JASA-ACS* and *TOMS* initiatives do not require use of a repository, many authors still provided access to their materials this way.

Version control systems are also widely used by researchers. Because AEA and *AJPS* authors are required to deposit in OpenICSPR and Dataverse, it is unclear how many of their authors would use version control systems otherwise. *IS* encourages authors to make materials available via Github, while *JASA-ACS*, *SC*, and *TOMS* do not.

A few authors used virtualization technologies to describe or encapsulate the computational environment. *IS* encourages authors to use Docker or ReproZip, the other initiatives do not. While not virtualization in the same sense, one *TOMS* author used a Python Virtual Environment to capture required dependencies.

PSEs are commonly used by social science and statistics researchers. In the computer science in mathematics cases, authors are generally developing software in other languages including Fortran, Python, C/C++, and Java.

Literate programming or reproducible document tools are less commonly used. While some authors provided Jupyter or R Markdown notebooks, only one provided a fully reproducible paper based on the LaTeX system.

Provenance capture tools are not widely used. The *IS* initiative encourages the use of ReproZip, which was used by the authors of two of the three replicated papers in that initiative.

[6] *SC* only requires use of archival repositories for consideration for the "Artifacts Available" badge.

7.5 Discussion

In Chapter 6, I concluded that the general factors contributing to computational transparency and reproducibility are similar across the seven initiatives. A few factors are closely related to the type of research being conducted and, while there is likely no one-size-fits-all policy, we can envision a comprehensive set of requirements that can be composed *à la carte* into policies by different communities. For example, systems research in computer science is likely to require more detailed information about the computational environment than the social sciences. Hardware versions and runtime states are more likely to effect the results of performance research than they are the results of statistical models and tests. However, reproducibility in both fields requires details about the software and operating system versions used. As discussed in Chapter 2, problems of numerical reliability and errors in third party software that may contribute to irreproducibility are not limited to any particular discipline. Just as we can envision a comprehensive set of requirements, we can envision a general model that can be used for the packaging and dissemination of research artifacts. In fact, several models exist today related to specific tools or communities. Research data archives, used by all seven initiatives, also implement general models that support a subset of initiative requirements.

In this section I also present a conceptual model of reproducible computational research artifacts. This model is intended to clarify the elements required for reproducibility, their relationships, and potential modes of representation. I also relate the model to the reproducibility infrastructure ecosystem to better understand how different tools relate to model elements. I propose a consolidation of multiple existing models via an extension of the research compendium concept first proposed by Gentleman and Temple Lang [105].

7.5.1 Elements of reproducible computational research artifacts

Table 7.4 presents a summary of the elements required for the packaging and dissemination of transparent and reproducible artifacts along with examples of their contents. In this section, I detail each of these elements followed by a discussion of their relationships. In this section, I use the term "package" to refer to the set of research artifacts required for assessment and verification.

Element	**Description and modes of representation**
Descriptive metadata	Information about the package including authors, organizations, funding agencies, tags, related identifiers. Since artifacts are associated with a single manuscript, much of this information can be shared/inherited. Descriptive metadata should included as structure metadata as part of the package or referenced externally (e.g., reference to manuscript).
Manifest	A file containing a list (name, description) of every object included in the final package. The manifest may be narrative text or structured and include citations or references to external resources.
Manuscript	The manuscript used to present the results may be included in the package as a reproducible document, as document source, or cited/reference externally via persistent identifier.
Computational workflow	Step-by-step instructions for reproducing reported results as 1) narrative text, 2) interactive notebooks, 3) wrapper scripts (e.g., a "master script"), or 4) other automated methods (e.g., workflow system).
Software	All software used in generation of results. Software may be included in the package as source/binaries, cited/referenced (including version and access date), or described in access protocol if non-public.
Data	All data used in generation of results. Data may be included in package, cited/referenced (including version and access date), or described in access protocol if non-public.
Logs	Execution output (e.g, stdout/stderr). Logs should include any output produced during the execution that led to the reported results and included directly in the package.
Results	Files containing results including data, figures, tables, etc. Results may be presented in the manuscript or as separate files used to produce figures, tables, etc in the manuscript.
Results provenance	Provenance information relating results to the code and data used to produce them specified. Provenance may be included as narrative text, via code comments/annotations, or captured automatically by an external tool.
Environment	Details of the computational environment. The environment may be described using one or more of narrative text, structured specification, or an image that is contained in the package, cited/referenced (with version and access date), or described in access protocol if non-public.
Resource requirements	Computational resource requirements (memory, cores, hardware (CPU/GPU), time) required to reproduce results. These requirements may be included as narrative text, structured specification, or captured during runtime.
Experimental context	Additional information about the experimental setup and workflow including subject selection, evaluation, and metrics. This may be included as narrative text or a (semi-) structured protocol.
Version	Since artifacts can change over time, version information must be included. Version information is meta-information about the package.
Certification	Certification statement or badge assigned as part of the review/verification process. Certifications are meta-information about the package likely not author-provided. The certification should link to externally published information including policy and review workflow documentation (including versions).
Review provenance	Information about the review/verification process including who performed the review, how long it took, errors found, policy/workflow version, resulting certification or badge. Review provenance is meta-information about the package and may be stored and referenced externally.
License and copyright	Code license, content copyright, and data copyright information. While not strictly required for reproducibility, license and copyright information are important for reusability.

Table 7.4: Elements of reproducible computational research artifacts

Descriptive metadata

The package should stand alone and provide sufficient descriptive information to be understood, reproduced, and related to externally published resources, such as the published paper. Because research artifacts are associated with a single publication, descriptive metadata (authors, title, description, organizations, funding sources, etc) will likely be identical. Related identifiers such as cited datasets, software, or environments should also be included (See Manifest above). Additional domain-specific metadata may be included.

Manifest

The term manifest is used here as in the context of a cargo manifest: a complete list and description of all items in the package. Every item should be listed in the manifest and no unnecessary files should be included. The manifest may include references to external resources (e.g., software, data, environment) via persistent identifiers.

Manuscript

The package must be associated with the draft or published manuscript. Under the "reproducible research" approach, the manuscript source is included as part of the package and can be easily rebuilt or recompiled whenever changes are made to any element of the package. This is more easily achieved with authoring systems such as LaTeX or markdown. In this case, the package must also include any software required to rebuild the manuscript. Another option is to reference the associated manuscript from the package via persistent identifier.

Computational workflow

The package must contain step-by-step instructions for reproducing reported results. The instructions may be provided narratively or, ideally, via some automation method. This includes interactive notebooks, a wrapper or master script, or the use of workflow automation systems.

Software

All software required for the reproduction of results must be provided either in the package; cited or referenced (including version and access date); or described via access protocol if non-public.

Data

All data required for the reproduction of results must be provided either in the package; cited or referenced (including version and access date); or described via access protocol if non-public. Source and analysis datasets must be provided. Intermediate datasets may be provided and are important if the process required to generate them is particular time or resource intensive (i.e., large-scale simulations).

Logs

Computational workflows typically produce informative output that can be captured in logs. Output often includes informative messages, warnings, and errors that may (or may not) related to the reproducibility of results. Capturing log output in the simple form of messages printed to standard output or standard error provide essential information for the review and verification process. Logs indicate which warnings and errors were present during the official run and whether they were ignored by the author. Log output from the execution that produced the reported results should be included in the package.

Results

Results reported in the manuscript must be provided either in the package or through reference to the manuscript itself. Ideally, the results are included either as data files or tables and figures generated by the provided scripts and workflow. Providing the results as part of the package allows for comparability during the assessment process.

Results Provenance

The relationship between results and the code and data used to generate them must be clearly specified. Results provenance can be specified via narrative or, more ideally, code/comments or annotations or the use of a provenance capture tool.

Environment

The package should provide details of the computational environment sufficient to reproduce the results. The environment may be described via narrative or, more ideally, through structured specification and virtual machine or container image. The environment may be contained in the package, cited/references (with version and access date) or described via an access protocol if non-public. The environment description includes all information that may impact reproducibility of results including software dependencies, compilers, operating systems, and hardware including versions and configuration settings. The environment description may include runtime state information, if relevant. The Whole Tale project has concluded that preserving the image (e.g., VM or container) ensures that the exact versions of installed software are available for the reproducibility assessment process as well as any future uses of the published package. The specification (or recipe) supplements the preserved image with the information that the author felt was most important to their work. However, images are typically quite large and resource-constrained repositories may have concerns about retaining complete environments. In this case the specification is acceptable, as long as it is recognized that they may not produce the exact environment.

Resource requirements

The package should include resources requirements required to complete the reproduction process including memory, cores, nodes, disk space requirements and estimated runtimes. Resource requirements can be provided narratively or through a structured specification (e.g., job submission script).

Experimental context

Some studies may require additional information about the experimental context. For example, the AEA requires authors to provide information about the complete experimental workflow including design summary, subject selection, data collection, etc. *SC* requires information about evaluation and metrics. Information about the experimental context may be include as part of the package or referenced externally.

Versions

Packages may be revised during the review and verification process or after a paper has been published. Because of this, it is important that the package be versioned and that the package retains this information.

Certification

Papers with packages that undergo review or verification are generally assigned some indicator such as a certification note or badge. The certification is important meta-information about both the paper and the verified version of the package. This information is generally not author-provided and is included as part of the pre-publication review workflow. Certifications should link to externally published information including associated policy and review workflow documentation. It should be possible for a future reader to identify certified papers and packages along with the processes that were used for certification.

Review provenance

The review and verification process results in the certification that provided package was used at a point in time to reproduce the results reported in the manuscript according to initiative policies and guidelines. The package version should include information about the verification including when it was conducted, who performed the review, how long it took, as well as any errors encountered. Ideally this information is made public, but there may be reasons to keep some of it private (e.g., protecting the identity of reviewers). However, this information proves useful in understand the operational factors that effect review times.

License and copyright

License and copyright information should be provided for all elements of the package. Although not strictly required for reproducibility assessment, license and copyright are important for future reusability. All packages should contain relevant code license, content copyright, and data copyright information.

7.5.2 Elements in context

Each of the elements described in the previous section contribute to computational transparency and reproducibility in different ways. Figure 7.1 illustrates the relationship between the various elements. Information about the computational environment and resource requirements are necessary to provision and configure (or access) the computational setup required for reproduction. The environment is necessary to execute the computational workflow, which requires software and data, and produces results and log outputs. The workflow defines the results provenance, which may be captured automatically using the same computational environment or specified manually. The results, logs, and provenance information can be used to compare the reproduction to the author's original outputs. If the author provides a reproducible document, the document creation process can also be used as a source of provenance. The figures and tables in the paper can be traced back to the workflow, software, and data used to create them. Other elements are not central to the reproduction, but are essential to understandability and reusability of the provided package. These elements are represented in the diagram as outside of the primary context for reproduction. Descriptive metadata is used primarily for discoverability and attribution. Accurate codebooks or related data documentation are necessary to understand variables used in the workflow. License and copyright information indicate how future researchers may reuse elements of the published package.

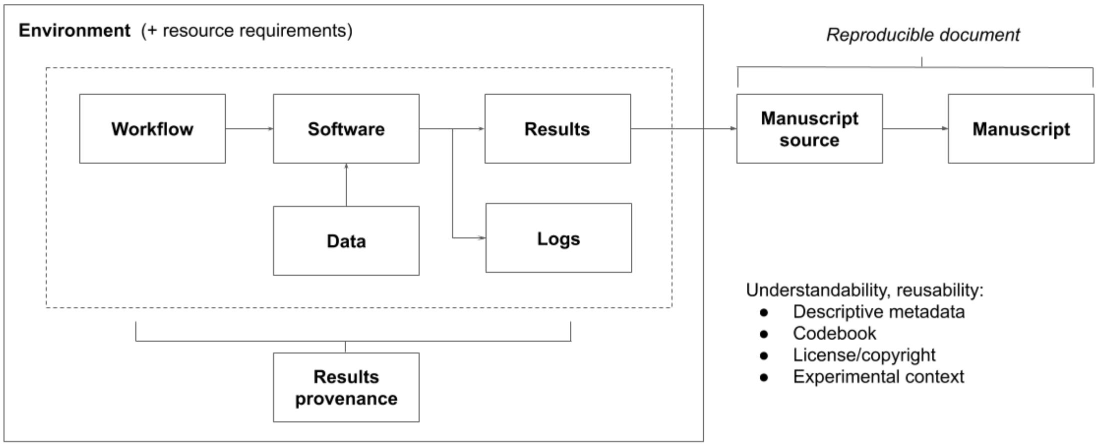


Figure 7.1: Elements in context: How research compendia elements relate to computational reproducibility.

7.5.3 Expanding the research compendium concept

Publishing computational research requires a variety of tools and infrastructure. Researchers have their own research environments or leverage shared computational infrastructure (e.g., campus, cloud, or national computing resources). They use different manuscript authoring systems. They may rely on version control systems for collaboration both on software and manuscripts. Editors and publishers rely on editorial submission and review systems, production publishing infrastructure, digital libraries, and research data archives. Research compendia, Research Objects, Binders, Capsules, SciUnits, ReproZips and Tales are all examples of packaging and dissemination formats that would need to fit into researcher, editorial, and publishing workflows and infrastructure to support the studied initiative's policies and workflows.

Each of these formats highlights different elements of computational transparency and reproducibility. Some focus on reproducible documents while others foreground computational workflows, capturing the computational environment, provenance information, as well as metadata and linked-data representations. Each also relates differently to the research infrastructure ecosystem. Some are designed to work with researcher tools and authoring environments; others with editorial and review workflows; or digital libraries and research data archives. To choose one format as the basis of this analysis would seem to foreground its strengths over those of another. Because Binders, Capsules, SciUnits, ReproZips and Tales all relate to specific platforms and implementations, they are not sufficiently general to serve as the basis of a conceptual model. Research compendia, as originally conceived by Gentleman and Temple Lang and extended by Marwick et al., are publication-centric and seem to imply the inclusion of the reproducible document. Research Objects are highly metadata-centric, focusing more on the technical representation via linked-data approaches than on what is being represented and why.

For the studied initiatives, two requirements stand out among the others. First, the packaged artifacts relate directly to the results reported in a single publication, whether or not the publication is included as a reproducible documented or referenced externally. The relationship between the packaged artifacts and publication are central. Second, the provided package must be assessable and/or verifiable through the expanded peer review process. This means that the package will likely undergo changes during the review process prior to being accepted or certified and, perhaps more importantly, pass through existing peer-review infrastructure.

While the concept of the Research Object is intuitive and compelling, its use implies a particular linked-data approach and seemingly foregrounds semantics over modeling. On the other hand, the research compendium concept, because of it's focus on the publication, seems better suited as a starting point for an expanded conceptualization. Marwick et al's notion of the compendium adds information about the computational environment as well as dissemination via version control systems and archival repositories. Ideally, these two different approaches will eventually merge.

Relationship to the manuscript

A research compendium corresponds directly to a single publication and is, in a sense, an extension of original publication. Because compendia will be assessed or verified, they are also part of the extension of the publication and review process. The compendium inherits many attributes from the publication including authors[7], title, description, funding agencies, etc. The compendium and manuscript may be versioned independently. Not all changes to the manuscript require changes to the compendium or vice versa. Ideally, compendia share metadata with the publication and are bi-directionally linked. Compendia may have additional metadata not present for the publication and some differences, such as a primary contact.

Relationship to editorial review

The compendium will be part of the peer review process. For the studied initiatives, packaged artifacts are typically not assessed until after the paper has been provisionally accepted. Authors are generally required to submit via conventional supplemental information mechanisms or via external systems. As such, draft compendia will benefit from the same handling as manuscripts, such as restricted access during review.

[7] It is possible that authors of the compendium and particularly the corresponding author could differ from the publication.

Relationship to archival repositories

Research compendia are well-suited to storage in archival repositories and many existing repository platforms already support some of the capabilities described above. Repository platforms support basic descriptive metadata and provide integration with journal review and publishing environments. However, compendia are often seen as objects deposited in repositories, not as central to the repository model. Today, Binders, SciUnits, ReproZips, and Tales are typically published as compressed zip files included in an archival record or package. A more powerful approach would be for the compendium to *be* the archival record or package. A reviewer or future researcher could download a self-contained version of the archival package that would represent the complete compendium including descriptive metadata and information about the assessment or verification process. Archivists and curators could operate on the research compendium both inside and outside of the repository platform where metadata and provenance information could be maintained.

Relationship to digital libraries

As an extension of the paper, compendia have many characteristics that are useful for discovery. For example, information about the programming languages used or details of the verification process (e.g., badges assigned) may be helpful for the identification of papers of interest to researchers. Today, there is generally a disconnect between digital library and archival discovery services. For example, metadata about artifacts published in archival repositories such as Dataverse, OpenICSPR, and Zenodo are rarely discoverable via publisher search engines. It is not possible, for example, to search within the Wiley system for *AJPS* articles that have undergone verification. The user must be knowledgable of both the digital library and repository systems.

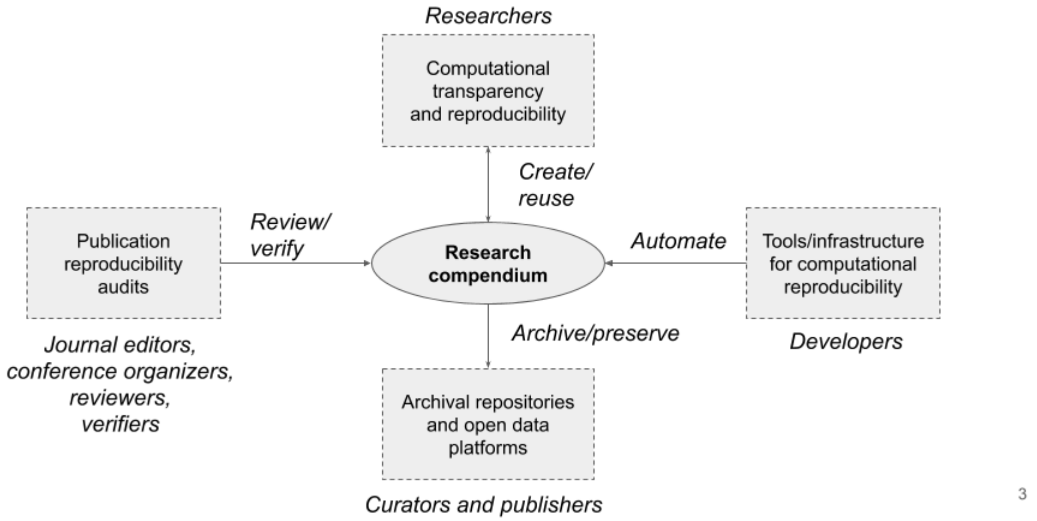

Researchers
Computational transparency and reproducibility
Create/ reuse
Review/ verify
Publication reproducibility audits
Research compendium
Automate
Tools/infrastructure for computational reproducibility
Journal editors, conference organizers, reviewers, verifiers
Archive/preserve
Developers
Archival repositories and open data platforms
Curators and publishers
3

Figure 7.2: Research compendium as a mediating information object informed by discipline-specific norms; reproducibility audit workflows; technical infrastructure; and archival repositories

Research compendium as a mediating information object

In this chapter I looked at the characteristics of tools and related packaging formats that enable computational transparency and reproducibility and considered how these characteristics relate to the assessment process and requirements of the seven initiatives. While the initiatives provide detailed guidance on the information required from authors, they provide limited direction on how artifacts should be packaged for review and dissemination. The primary tool required by the initiatives is the research data repository, largely because of the commitment to long-term archiving, issuing of persistent identifiers, and in some cases existing integrations with editorial and publishing processes. Virtualization and provenance capture technologies have had limited adoption across the seven initiatives, but could possibly simplify the process of packaging and review. Nascent metadata standards, such as RO-Crate [240], present the best option for the description, encapsulation, and dissemination of complex research artifacts.

Chapter 8

Conclusions

Many research communities concerned with the rigor or trustworthiness of computational results are exploring approaches to improving or ensuring the reproducibility of published research. A general assumption, as evidenced by the studied initiatives, has been that the solution is to expand the peer review process to include the assessment of computational artifacts and results. The research presented in this dissertation aims to further our understanding of how some communities are leveraging peer review to ensure computational reproducibility as well as the social and technological challenges that they face. This is achieved through the investigation of the following research questions:

- RQ1. How are computational transparency and computational reproducibility operationalized through publication reproducibility audits?
- RQ2. What are the characteristics of research artifacts that make them computationally reproducible (or irreproducible)?
- RQ3. What are the characteristics of tools and packaging formats that enable computational transparency and reproducibility?

In this chapter, I summarize the key findings followed by a discussion of the implications of this research and future research directions.

8.1 Summary of key findings

RQ1: Computational transparency and reproducibility are operationalized through policies and review workflows, which are affected by both social and technical factors. Organizationally, all of the initiatives have introduced specific new editorial roles responsible for the development and enactment of new policies. The policies define the information required of authors and criteria for assessment, while other key operational decisions determine when the review is conducted, what is assessed, and by whom. In a remarkable similarity, in all seven initiatives the review process is conducted post-acceptance, such that the reproducibility review has no material effect on the acceptance decision. This is likely to encourage community buy-in while minimizing the impact of the new policies on the publication review process. The initiatives differ with respect to policy mandates, which determine whether the assessment process is uniformly applied. Opt-in policies reduce the risk of push-back or a negative impact on submissions, allowing initiatives to scale up as demand increases within the community, but also risk selectivity-bias as participating authors are already confident in the quality of their work. Additional central differences can be seen in what is reproduced and by whom. The initiatives vary from the assessment of materials only to full reproductions of results. They also rely on reviewers with different degrees of expertise, including undergraduate students, graduate students, practitioners and peers.

RQ2: Core factors impacting computational reproducibility are shared across the initiatives, but vary by type of research. I identify four core factors impacting computational reproducibility including: complete documentation of the computational workflow; accessibility of precise versions of software and data used in the generation of results; sufficient information about the computational environment to enable reproduction; and long-term accessibility of research artifacts. While initiatives policies may include other requirements, these four factors are essential to enabling reproduction and reproducibility assessment. How materials are made accessible depends on whether the research relies on private or protected resources (e.g., data, software, or hardware). In this case, authors cannot provide the materials directly but can be required to provide detailed access protocols describing how someone with appropriate permissions can gain access. Private resources are no longer an exclusion, but may impede the reproducibility assessment process. What constitutes "sufficient information about the computational environment" varies by the type of research being conducted. Configuration and versions of operating systems, dependent software, hardware or even networks may play an important role. In some research areas, system runtime state may also be an important factor in reproducibility. The depth of information required to describe the "environment" differs across these cases, but is similar across different types of research. A political scientist conducting research using a high-performance computing environment will likely require different types of information than a colleague working on their laptop.

RQ3: There are material gaps in existing and available infrastructure. Returning to the four core factors discussed above, many technical solutions exist to support authors in the dissemination of reproducible computational research artifacts. Scientific workflow automation systems, reproducible documents, automated provenance capture, and virtualization technologies can be used to provide detailed information about the computational workflow, software, data, and computational environment used in the generation of results. However, these tools are largely unused in the current initiatives. Currently, initiatives are focused on technical infrastructure required for reproducibility assessment. The editorial and publishing infrastructure central to the manuscript peer review process are not well-suited for reproducibility review. As a result, several initiatives have developed custom infrastructure to support the tracking and review process. Additionally, initiatives face challenges in gaining access to the computational resources and licenses required to conduct reproductions.

8.2 Generalizability of findings

I believe that the findings from this investigation have broader implications beyond the studied initiatives, despite the common critique that case study research is not generalizable. As in most case study research, this investigation suffers from a small sample size. However, the qualitative case study approach has allowed me to analyze in depth the individual cases and the many complexities that underlie their implementations. The specific cases were selected because they are representative of similar initiatives within their respective fields. As such, the basic framework for analysis and reported findings should be transferable to other initiatives with similar organization types.

Any new or existing reproducibility initiative will have the same key characteristics as those identified in this study. They must define an organizational structure for reproducibility review; determine who will conduct the review, at what depth, and how they are incentivized. They must establish a policy and mandate and make specific decisions about review workflows and infrastructures that will be shaped by their current organization. While the analysis of more reproducibility initiatives may expand our understanding of some of these characteristics, the basic framework for analysis can still be applied

Similarly, the framework and findings described in Chapter 6 for the analysis of the reproducibility of research artifacts is also transferable to other initiatives. A central finding of this investigation is that, despite the broad differences between the selected cases, the factors involved in the assessment of computational reproducibility are remarkably similar. Not only is the framework for analysis applicable to initiatives that are not part of this study, but I expect that the identified core factors involved computational reproducibility apply beyond the studied cases.

The normative guidelines presented in Chapter 10 reflect my belief that the findings from this study can be applied to new reproducibility initiatives across the sciences. These guidelines are intended to support journal and conference leadership; tools and infrastructure developers; and funding bodies to better understand how decisions about the operationalization of the reproducibility review process impact our understanding of what is meant by 'reproducibility.'

8.3 Next Steps

While promising, these initiatives and associated infrastructure advancements have not yet been proven to have desired effects on research quality or trustworthiness. I conclude this dissertation with proposals for possible next steps based on the results of this study. First, further efforts should be made to study whether the initiative policies and associated tools are having the desired effects on research quality before advocating for widespread adoption. Second, further studies should investigate whether advancements in reproducibility tools and infrastructure 1) improve the long-term reproducibility of computational research artifacts and 2) reduce the burden on authors, reviewers, and related stakeholders in the audit process. Initiatives such as those currently implemented by AEA or APSA journals present ideal opportunities to study both the effects of reproducibility audits on research quality and the potential effects of tools and infrastructure advancements on the audit process. Both communities have an extensive history of implementing and evaluating policy changes intended to improve reproducibility and replicability of published research. The AEA initiative is ambitious, open, and largely intended to inform other initiatives within the social sciences. As academic societies with centralized reproducibility assessment operations, it presents a unique opportunity to address both proposed studies.

The AEA initiative presents a unique opportunity in terms of both scale and mandate. Funding efforts to measure the effect of different approaches in cooperation with the AEA would provide maximum benefit. It seems likely that authors will minimize the effort they put into packaging their work for review (hence the high-frequency of errors). Infrastructure that supports the verification of computational reproducibility should therefore minimize both 1) author effort in complying with journal policies and 2) reviewer time required while also 3) maximizing the long-term technical reproducibility of the provided artifacts. Authors will only adopt a new tool or convention if the value exceeds the cost of learning it. Similarly, reviewers are unlikely to spend time learning new tools unless they simplify the review process. It will help if tools are accepted community wide.

The focus on computational reproducibility, while seemingly pragmatic, may prove counterproductive in the long run. The true measures of success of these and related initiatives is whether they impact the quality and trustworthiness of reported results. The ultimate measure may be the rate of true replications enabled by these initiatives. Following the example of Berry et al. [32], we might expect higher rates of true replications – and successful replications – when studies are based on earlier work with verified reproducible artifacts.

Chapter 9

Epilogue

Reproducibility[1], it is often argued, is a cornerstone of science and central to the process of establishing scientific facts. “Non-reproducible single occurrences,” the philosopher Karl Popper once noted, “are of no significance to science” [220]. However, there is a well-recognized paradox in many scientific fields today. While published studies are expected to be reproducible, most scientists do not reproduce the results of others or read about reproductions [43, 90, 187]. Journals are also unlikely to accept exact reproductions unless they are part of extensions that provide new information. As noted by Casadevall & Fang [43], this bedrock assumption of science is rarely tested and generally lies in our trust in the ability and integrity of individual researchers and systems.

9.1 Trust but Verify

Trust is essential to science and scientific knowledge production [118, 274]. Progress in research depends on the willingness of others to believe and accept new knowledge as reliable and trustworthy. Sources of epistemic trust include trust in methods as well as trust in individuals [68, 274]. The scholarly communications process determines the amount and types of information that researchers are required to provide for their results to be assessed and deemed trustworthy. The process of peer review has over time become central to the determination of trust in scientific results and remains one of the most important factors for determining the quality and trustworthiness of research today [261]. Through this process, reviewers assert their confidence in the impact as well as potential reproducibility, replicabil-

[1] In the broadest sense of the word.

ity, and trustworthiness of reported results – generally without undertaking a reproduction or replication themselves.

To address concerns about the trustworthiness of the results of computational research, the seven initiatives investigated in this study have each made changes to the peer review process and in each case increased the information required of authors in order for their work to be published. These initiatives suggest that there is a distinct value to science for researchers who leverage computational and data-driven methods to adhere to practices that ensure reproducibility of results and to make available the materials that underlie claims for assessment. However, there is a duality to these efforts. On one hand, they are concerned with changing researcher practice to improve the quality and trustworthiness of computational research while, on the other, they are concerned with protecting the integrity of published research independent of researcher practice.

By expanding the peer review process, these initiatives provide new incentives for authors to disseminate reproducible computational research artifacts. This can be seen as an attempt to correct a misalignment in current incentive structures. If researchers will not voluntarily adopt these practices, then the mechanisms of peer review and pre-publication assessment can be used to enforce some degree of transparency and reproducibility in their work.

For communities interested in adopting similar policies and practices, there is no blueprint, but the studied initiatives can serve as examples. They have each faced challenges in community readiness; gaps in social and technical infrastructure; and made different decisions concerning the operationalization of review workflows. The results of this study demonstrate a high degree of overlap in policies and requirements of authors and suggest an opportunity for the development of general policies and technical requirements for dissemination.

This study has presented an in-depth exploration of the seven initiatives. In Chapter 5, I considered the question of why and how each initiative operationalizes the reproducibility assessment process. In Chapter 6, I explored the characteristics of research artifacts that contribute to computational reproducibility. Finally, in Chapter 7, I looked at how different tools, infrastructure, and packaging formats enable computational reproducibility and may aid the reproducibility assessment process. In this chapter, I consider the implications of my findings for research communities considering the adoption of the NASEM recommendations as well as tool and infrastructure developers.

9.1.1 The Duality of Computational Reproducibility

Broadly speaking, efforts encouraging or requiring reproducible computational research fall into two categories: those concerned with improving research quality and trustworthiness through changing researcher practice and those concerned with protecting the integrity of research at the point of publication.

Many have argued that computational reproducibility should be the standard for all computational work across the sciences [81, 157, 213, 252]. It serves the interest of the researcher, their collaborators, and future readers. However, due to the misalignment of incentive structures discussed below, many researchers do not adhere to these practices voluntarily. Research outlets, such as those at the center of this study, enforce computational reproducibility independent of individual researcher practice. While these initiatives may hope to influence practice, they do not take on the burden of enforcement for the benefit of the author. They do it to protect their identity and the integrity of the research that they publish.

This distinction underlies Peng [209] and Leek & Peng's [161] critiques of reproducible computational research. If the goal is to increase the quality and trustworthiness of computational research, then enforcement at the point of publication is too late. They argue that efforts would be better spent improving researcher education – the "preventative" approach. However, outlets concerned with the quality of computational research published today cannot wait for the effects of broad-base educational initiatives. They instead leverage existing incentive and enforcement mechanisms through the expansion of the peer review process.

9.1.2 Correcting Misaligned Incentive Structures

Advocates for computational transparency and computational reproducibility have recognized that these activities require significant changes to community norms and academic incentive structures [252]. For researchers, preparing reproducible computational artifacts for dissemination requires effort that is often unrewarded and takes away time that can be spent on new activities. The studied reproducibility initiatives are attempting to change this equation by encouraging or requiring authors to provision reproducible artifacts. In doing so, they have had to also address the incentives for their own reproducibility reviews.

As discussed in Chapter 2, economists have been studying the incentive structures related to computational reproducibility and replicability for decades. In their study of the absence of replications in empirical economics research, Feigenbaum & Levy [90] describe the powerful disincentives for authors to share high-quality materials and conclude that journals would need to increase the information required of authors for materials to be made available. Mirowski & Sklivas [187] suggest that increasing true replications might require paying replicators or requiring apprentice researchers to attempt replications in the course of their training.

Even today, evidence suggests that only a small number of researchers will voluntarily provide computationally reproducible research artifacts. Table 2.1 looks at the rates of voluntary participation within the SIGMOD initiative, which was relatively high when the initiative began (70% of accepted papers participated) but steadily decreased over time to less than 10%. In the current study, the voluntary initiatives (Biostatistics, TOMs) have had remarkably low participation rates with fewer than 5 over a multi-year period. Those with mandatory policies have necessarily had higher compliance rates, but have also had to address the high-cost of the reproducibility assessment process.

Incentives for conducting the reproducibility assessment differ depending on how the review process is operationalized. Full reproductions require considerable effort and create an additional burden for the average peer reviewer. Increasing the workload on traditional reviewers to include reproducibility assessment risks additional "reviewer fatigue" [39, 249]. Initiatives have addressed this in one of three ways. They have 1) engaged students or paid professionals to undertake full reproductions, 2) minimized the information required for review (e.g., appendices), or 3) implemented opt-in policies to limit the number of reviews. This is likely because there is limited value to peers in the systematic reproduction of all accepted research. Further engagement of students or apprentice researchers may provide a solution to the incentive problem, but comes with additional risks. They may be more susceptible to bias or risk damaging future careers by challenging the work of more senior scientists.

All seven of the studied cases are exploring ways to change norms and incentive structures for their researchers and communities. It is the policies adopted by journals and conferences and changes to peer review that appear to be having the largest impact. As has been discussed previously, changes to policies work within current incentive structures and provide a needed forcing function. The journals and conferences are incentivized to require authors to provide additional information and to expand the review process in order to protect their own integrity and identity. Authors are incentivized to provide required materials in order to have their work published.

9.1.3 Factors in Community Readiness

NASEM recommendation 6-4 states that journals should "consider ways to ensure computational reproducibility" while acknowledging "technological and practical challenges." In addition to the incentive structures discussed above, any new initiative attempting to implement this recommendation should understand both the social and technological factors in community readiness.

The studied initiatives represent the latest developments in decades-long community efforts to address concerns about computational research quality and trustworthiness. While recent policy changes may be motivated by the “reproducibility crisis” narrative, they are largely made possible by groundwork already laid within each community, including both social and technical infrastructure.

The success of these initiatives is due in part to substantial “cultural inertia” and a ready “installed base” [247]. Members of many of these communities have long debated the merits of publishing reproducible computational research [66, 81, 145, 154, 214]. They have adopted previous policies related to software review [93] as well as code and data availability [10, 30, 184, 276] and rely on advancements in related technical infrastructure, including research data repositories. Five of the initiatives (AEA, AJPS, JASA-ACS, SC, TOMS) have coincided with broader community and association efforts including changes to ethics guidelines with respect to research transparency and reproducibility.

9.1.4 Gaps in Infrastructure

In recommendation 6-3, the NASEM report suggests that funding agencies consider further investment in the development of tools and infrastructure in support of computational reproducibility as well as training and outreach for researchers to leverage them in their work. As detailed in Chapter 7, there are a wide variety of existing tools designed to improve or simplify the process of conducting and packaging reproducible computational research artifacts, but they are largely unused in the studied initiatives. The most widely used tools include editorial and publishing infrastructure, archival repositories, and computational infrastructure for re-execution.

In the context of publication reproducibility audits, editorial and publishing infrastructure is central. As detailed in Chapter 5, the requirement of authors to submit computational artifacts and the introduction of new review roles and workflows has exposed gaps in current editorial infrastructure. Widely used tools such as Editorial Manager and ScholarOne lack many of the capabilities required for the review and verification of computational materials. They lack facilities to manage and review data and code and assume that reviewers are part of the standard peer review process. As a result, several initiatives have developed custom tools and workflows to manage review, tracking, reporting, and author communications. This suggests an opportunity to invest in new infrastructure that incorporates the lessons learned from these initiatives.

All seven of the studied initiatives rely heavily on existing archival infrastructure. Research data repositories including Dataverse, OpenICSPR, Zenodo, and Mendeley Data are central to initiative policies and also voluntarily used by authors. These mature repositories provide storage and assurances for long-term archiving and preservation along with issuing persistent identifiers required for linking to publications. They reduce burden on authors, associations, and publishers for maintaining long-term access to published materials. While some repositories include features for integration with editorial and publishing workflows, they generally lack capabilities required for assessing computational reproducibility. Per the related recommendation 6-5, funding agencies should continue investment in the expansion of capabilities in these existing platforms.

Initiatives that perform full reproductions rely on the availability of computational resources provided by reviewer host institutions. In the cases of the AEA and AJPS, these resources are currently provided Cornell and UNC respectively. Reviewers and verifiers have access to institutional resources including virtual machines, batch compute clusters, licensed software as well as technical support resources. The other initiatives rely on host institutions for individual reviewers and provide no centralized resources. There is an opportunity to leverage existing investment in national research computing infrastructure in support of the reproducibility assessment. For example, NSF Jetstream and XSEDE could provide managed access to many of the same resources used by researchers for the review and assessment process.

Many of the specialized tools developed for computational reproducibility (see Chapter 7) have the potential to reduce author or reviewer burden while also improving computational reproducibility. Scientific workflow platforms, automated provenance capture systems, as well as container and virtualization technologies simplify the complex processes of workflow re-execution and capturing details of the computational environment. It is notable that many of these tools are not widely used as part of current initiatives nor have their effect on either reproducibility or efficiency been studied. This suggests an opportunity for funding agencies to support research into the application of these new technologies in the context of current audit initiatives. For example, it may be possible to study whether the use of these tools may reduce burden on authors or reviewers in cooperation with initiatives, such as the AEA or AJPS.

9.1.5 Opportunities for General Policies and Packaging Formats

In recommendation 4-1, the NASEM report details the information required of authors to ensure the reproducibility of computational results. As detailed in Chapters 5, 6, and 7, all seven of the studied initiatives have highly overlapping policies and requirements for authors. Key differences can be seen when addressing private or proprietary resources; computational scale and complexity; and what constitutes the computational environment. While the depth of information required for individual elements may differ, it should be possible to create a general set of standard guidelines that can be used to compose community- or journal-specific policies and checklists.

As discussed in Chapter 6, accessibility of data, software, and hardware resources is essential to the reproduction process. However, there are legitimate cases for restricting access to each. For example, human subjects information and proprietary software and hardware may limit who is able to access resources required for reproduction. The AEA initiative introduces the concept of the “access protocol” for protected information. This same notion is applicable for other types of resources. In the absence of direct access, researchers can provide detailed information about how an authorized individual can gain access.

As evident in the original *JMCB* study, all disciplines are likely to face issues of computational scale. Large-scale modeling on state-of-the-art computational resources can be found in all areas of research and will face the same challenges when assessing reproducibility. Initiatives such as the AEA and AJPS will benefit from the experience and expertise of the IS, TOMS, and SC initiatives when considering how to approach the assessment of the results of studies where full reproduction is impractical or impossible.

Details of the computational environment are also central to the reproduction process, but the level of detail required depends on the nature of the research and the types of resources required. In general, reproducing the results of small-scale computational processes may require only information about the operating system, common PSEs and related software dependencies. Systems performance research may require additional information about compiler, hardware, and network configurations and even runtime state information. While it is unlikely that statistical hypothesis tests in the social sciences will ever need runtime state information, it is likely that performance research across computer science and engineering would benefit from am common set of requirements.

The requirements outlined in NASEM recommendation 4-1 are broadly applicable. Access to data, computational workflow and computational environment are common to all seven initiatives. Where these differ are in the details, which upon inspection also retain much commonality.

9.1.6 Reproducibility of What, By Whom, What is Gained?

In Chapter 3, I use Radder's typology [225] and the PRIMAD model [95] to better understand the dimensions of computational reproducibility. These frameworks are the basis for questions explored in Chapter 5 concerning *what* is being reproduced by *whom* and what information is gained through each of the studied initiatives. The seven cases present a variety of different approaches to computational reproducibility along these three dimensions.

With respect to *what* is being reproduced and by *whom*, the seven initiatives adopt four different approaches. These include assessments of reproducibility without reproduction, partial reproductions, full reproductions, and full reproductions with extensions. The reproducibility assessment process is undertaken by individuals with a wide range of experience and expertise including peers in the community, expert practitioners, as well as advanced graduate and even undergraduate students. In all cases, technical and methods knowledge appear to be more important than theoretical knowledge. The traditional peer review process remains central to determining the theoretical contribution of the work while the reproducibility review determines only whether materials are effectively made available to reproduce reported results. In this sense, all but one case (IS) are concerned with Radder's "material realization," where theoretical knowledge is unnecessary, but tacit or technical knowledge is.

The PRIMAD model was proposed to capture the dimensions of scientific reproducibility in computing sciences. This is a much broader notion of reproducibility than is presented in the NASEM report's definition of "computational reproducibility." The PRIMAD model can help clarify the tension between this narrow notion of "computational reproducibility" and scientific reproducibility in computational sciences. Through the lens of the PRIMAD model, we can see that six of the seven initiatives are concerned only with assessment and verification. At most, the actor (A) and possibly platform (P) are changed – although the change in platform is not necessarily even seen as intentional. A key objective of PRIMAD is to understand the "information gained from different types of reproducibility" activities. In the context of PRIMAD, little is gained from these initiatives in terms of new scientific knowledge other than the confidence that the authors have provided the required information. The IS initiative is the only initiative that has adopted other elements of the PRIMAD model, as invited reviewers are asked to extend the reported research in some way.

9.2 Experiments in Scholarly Publication

The *JMCB* study was an experiment designed to measure the effect of a journal policy change on the availability of data and code associated with published research. It is, of course, quite unusual for journal editors to approach policy changes as studies themselves. They are more likely undertaken in response to problems identified by the research community.

Each of the seven initiatives can be seen as a potential experiment in how changes to incentive structures and information requirements of authors impact the availability of materials and the quality, reproducibility, and trustworthiness of published research. Unfortunately, they are not designed as such. To do so would require establishing a set of hypotheses and measures. In the *JMCB* study, the hypothesis was that the journal policy change would increase the availability and quality of replication materials. Metrics included the number of papers that provided materials before and after the change, before and after acceptance for publication, and replication rates of provided materials.

It is argued but rarely studied that computational reproducibility practices may increase the impact of both papers and journals. In a small study of citation rates of papers in signal processing research, Vandewalle [264] suggests that code sharing is associated with increased citation rates. Anecdotally, editors of three journals in the current study also suggested that an impact factor increase correlated with the adoption of new policies.

The NASEM report recommendation 6-4 states that journals should "consider ways to ensure computational reproducibility for publications that make claims based on computations, to the extent ethically and legally possible." If the goal is to change researcher norms and to increase the quality and rigor of published research in the interest of reproducibility and replicability, journal reproducibility audits are one approach. Certainly mandatory policies and audit processes will increase the availability of materials in conformance with policies. However, it is still unclear whether they have the desired effects of either changing researcher behavior or increasing the overall quality of research. Current initiatives represent an ideal opportunity to study whether audit processes are worth the cost[2].

[2] [87] reports $180 per article for graduate student reproduction, which is consistent with information provided by study participants

Chapter 10

Recommendations for New Initiatives

I close this dissertation with a set of recommendations for communities considering the adoption of policies and workflows for the verification and dissemination of computational research artifacts for transparency and reproducibility. As discussed throughout this dissertation, the studied initiatives build on foundations established within their communities for years and decades prior to the adoption of new policies. I also caution that the effects of these initiatives are still not well understood. While the adoption of reproducibility review and audit processes seems sensible, whether they improve the quality and trustworthiness of published research remains to be seen.

A goal of this study is to develop a set of normative guidelines and recommendations for communities interested in pursuing the development of similar initiatives. Here I present these recommendations targeting journal editors, conference organizers, and funding agencies.

10.1 Initiatives and Policies

1. **Assess community readiness and commitment.** While requiring computational reproducibility may seem like a natural step, it imposes substantial changes on researcher, journal, and reviewer workflows as well as associated infrastructure. Making this type of change will require significant investment of time, leadership, and a commitment to carry through the vision. Consider the impact of a policy that is later abandoned due to changes in leadership. Assessment of community readiness or commitment can be conducted informally among publication leadership, through community surveys (e.g., [91]), or symposia dedicated to the discussion of policy changes (e.g., *PS: Political Science & Politics* 28:3 and *Biostatistics* 11:3 discussed below). These

open discussions can provide a diverse set of viewpoints and a sense of how the broader community will react to policy changes.

2. **Identify motivations**. Determine whether your community is more concerned with the reproducibility of results or software reuse and trustworthiness (or both) in the development of policies and review workflows. Policies that encourage the development of re-usable research software do not necessarily require reproducibility of scientific results. Similarly, policies that enforce the reproducibility of results do not necessarily require or even encourage the development of re-usable software.

3. **Reproducibility of what, by whom?** Operationalizing the assessment process will require you determine whether you intend to undertake reproductions (e.g., re-execute code and compare results) or assess potential reproducibility (e.g., review provided materials without actually reproducing results). Additionally, you will need to determine who conducts the review (e.g., peers, expert practitioners, advanced graduate students) and how they are incentivized to participate (e.g., editorial role, financial compensation, goodwill). This will determine or be determined by how important different types of expertise – technical or theoretical – are deemed important in the assessment process as well as the availability of financial or other incentives.

4. **Publish policies and publication workflows**. Develop a clearly written public policy with both author and reviewer guidelines including provisions for private or confidential resources and handling instances of non-reproducibility. All policies, guidelines, and workflow documents should be published with persistent identifiers and versioned. Ideally, all manuscripts and artifacts reviewed under a particular policy version are linked to that version.

5. **Recognize infrastructure gaps**. Identify all existing technical infrastructure used by your journal or conference and understand the gaps. Policies and workflows may be constrained by infrastructure including publisher and editorial management systems, digital libraries, and repositories. Many of the cases presented in this study have developed custom tools and workflows to address limitations in existing infrastructure. Understand the availability of computational resources and licenses required for re-execution.

6. **Address the "big" and "private"**. Include provisions for how to handle research conducted using large-scale or private computational or data resources. Consider the

use of access protocols (i.e., documented steps to gain access to resources, even if difficult) and reduction tests to assess reproducibility where reproduction is difficult.

7. **Leverage existing repository infrastructure.** Fortunately, mature archival repository infrastructure exists today and in some cases already has in place some features to support reproducibility assessment. Platforms such as Dataverse, OpenICSPR and Zenodo are widely used among existing initiatives.

8. **Have "the talk" (or "talks").** Special issues or symposia dedicated to the discussion of policy changes present the opportunity to formally collect community input. Examples can be found in *PS: Political Science & Politics* 28:3 and *Biostatistics* 11:3. These discussions are also invaluable for other communities considering the adoption of similar policies and practices.

9. **Instrument the review process and measure policy effect.** There are many ways to measure the effect of policy changes. Operational metrics include the number of papers assessed under a given policy, amount of time required to conduct the assessment, amount of time added to the publication process, number of resubmissions, and the type and magnitude of errors. By instrumenting the review process and publishing anonymized data, it may be possible to assess the effect of your and similar policies on the research and publication process.

10.2 Information Standards and Packaging Formats

Table 10.1 presents a summary of recommended guidelines to include in computational reproducibility policies. I propose that a general set of guidelines could be developed by a multi-disciplinary body to reduce the burden of new initiatives in devising comprehensive policies and communicate broadly effective strategies across communities.

Factor	**Description**
Manifest	Name and description of all artifacts required to complete the reproduction process including all files in the "package" and references to any external artifacts (e.g., referenced data, software, workflows, protocols, etc).
Workflow	Complete computational workflow must be clearly described either narratively, through a master script, or other workflow submission protocol. For experimental work, the documented workflow may require information about non-computational steps.
Software	All author-developed software required to complete the reproduction process must be accessible either directly or through a well-defined access protocol.
Data	All original and externally references data required to complete the reproduction process must be accessible either directly, through citation (including version), or through a well-defined access protocol. This includes any source data used to generate analysis datasets used in the final results. All data must be accompanied by a codebook or similar documentation describing variables with accurate labels.
Environment	The complete computational environment required for reproduction must be accessible or well-documented. This includes details about any dependent software, operating system, compilers, or hardware details that may impact the results, including specific versions and settings for each. Required resources (disk space, memory, cores, running time) and runtime state (single user, hot/cold cache, process pinning) may also be required. Detailed installation and configuration documentation, an image of the environment (e.g., VM or Docker) or protocol describing how to gain access to an existing system are required for the reproduction process.
Experimental context	The discipline-specific experimental context including experimental design, subject selection, evaluation methods, and metrics.
Results	For reproducibility assessment, the primary results must be provided for comparison. Results may be in the form of references to tables, figures, and in-text analytical claims in the associated manuscript or as separately packaged artifacts (e.g., images or data files).
Provenance	The relationship between all code and data to the results must be provided. Provenance may be described narratively, in the manifest, or through comments/annotations in the source code.
Trustworthiness	In the absence of the ability to conduct a full reproduction, information to understand steps taken by authors to ensure the validity of their work.
Publication	The complete set of artifacts (including access protocols for private or proprietary resources) should be published to an appropriate archival repository. Links should be made between the published paper and associated artifacts. Ideally, both the paper and published artifacts include information about the assessment process, such as links to the specific version of the policy and guidelines used for evaluation.

Table 10.1: Summary of recommended guidelines for computational reproducibility initiative policies

Bibliography

[1] M. Altman, L. Andreev, M. Diggory, G. King, A. Sone, S. Verba, and D. L. Kiskis. A Digital Library for the Dissemination and Replication of Quantitative Social Science Research: The Virtual Data Center. *Social Science Computer Review*, 19(4):458–470, Nov. 2001.

[2] R. M. Alvarez, E. M. Key, and L. Núñez. Research Replication: Practical Considerations. *PS: Political Science & Politics*, 51(2):422–426, Apr. 2018.

[3] R. G. Anderson and W. G. Dewald. 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,*, November/December 1994:79–83, 1994.

[4] Anonymous. Prologue. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 16(2):88–88, 1967.

[5] Anonymous. Algorithms Section: Introduction. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(3):381, 1997.

[6] Anonymous. Enhancing reproducibility. *Nature Methods*, 10(5):367–367, May 2013.

[7] Anonymous. SC15 Paper Highlighted in Effort to Raise HPC Research Integrity to Serve as Basis for SC16 SCC Reproducibility Application Challenge, Sept. 2016.

[8] APSA. *A Guide to Professional Ethics in Political Science*. American Political Science Association, second edition, 2012.

[9] S. Aseeri, B. K. Muite, and D. Takahashi. Reproducibility in Benchmarking Parallel Fast Fourier Transform Based Applications. In *Companion of the 2019 ACM/SPEC International Conference on Performance Engineering*, ICPE '19, pages 5–8. Association for Computing Machinery, 2019. event-place: Mumbai, India.

[10] O. Ashenfelter, R. H. Haveman, J. G. Riley, and J. T. Taylor. Editorial Statement. *The American Economic Review*, 76(4):v–v, 1986.

[11] A. S. Association. *Ethical Guidelines for Statistical Practices*. American Statistical Association, 2016.

[12] A. S. Association. *Recommendations to Funding Agencies for Supporting Reproducible Research*. American Statistical Association, Jan. 2017.

[13] K. Baggerly. Disclose all data in publications. *Nature*, 467(7314):401–401, 2010.

[14] K. A. Baggerly and D. A. Berry. Reproducible Research. *AMSTATNEWS*, Jan. 2011.

[15] K. A. Baggerly and K. R. Coombes. DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY. *The Annals of Applied Statistics*, 3(4):1309–1334, 2009.

[16] K. A. Baggerly, J. S. Morris, and K. R. Coombes. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20(5):777–785, Mar. 2004.

[17] D. Bailey. Misleading performance in the supercomputing field. In *Supercomputing '92:Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 155–158, Nov. 1992.

[18] D. H. Bailey. Twelve ways to fool the masses in performance evaluation. *Supercomputing Review*, pages 54–55, 1991.

[19] D. H. Bailey, J. M. Borwein, and V. Stodden. *Facilitating Reproducibility in Scientific Computing: Principles and Practice*, pages 205–231. John Wiley & Sons, Ltd, 2016.

[20] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, May 2016.

[21] L. A. Barba. Terminologies for Reproducible Research. *arXiv:1802.03311 [cs]*, Feb. 2018. arXiv: 1802.03311.

[22] M. Barni and F. Perez-Gonzalez. Pushing science into signal processing [my turn]. *IEEE Signal Processing Magazine*, 22(4):120–119, July 2005.

[23] B. Baumer and D. Udwin. R Markdown. *WIREs Computational Statistics*, 7(3):167–177, 2015.

[24] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, and et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, Feb. 2013.

[25] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*, pages 1–1, July 2010.

[26] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*, July 2010.

[27] R. Becker and J. Chambers. Design and Implementation of the S System for Interactive Data Analysis. In *The IEEE Computer Society's Second International Computer Software and Applications Conference, 1978. COMPSAC '78.*, pages 626–629, Nov. 1978.

[28] C. G. Begley and L. M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, Mar. 2012.

[29] A. Bennett, A. Barth, and K. R. Rutherford. Do We Preach What We Practice? A Survey of Methods in Political Science Journals and Curricula. *PS: Political Science & Politics*, 36(3):373–378, July 2003.

[30] B. S. Bernanke. Editorial Statement. *The American Economic Review*, 94(1):404–404, 2004.

[31] D. Bernstein. Containers and Cloud: From LXC to Docker to Kubernetes. *IEEE Cloud Computing*, 1(3):81–84, Sept. 2014.

[32] J. Berry, L. C. Coffman, D. Hanley, R. Gihleb, and A. J. Wilson. Assessing the Rate of Replication in Economics. *American Economic Review*, 107(5):27–31, May 2017.

[33] K. Bollen, J. T. Cacioppo, R. M. Kaplan, J. A. Krosnik, and J. L. Olds. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science: Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*. National Science Foundation, 2015.

[34] P. Bonnet, S. Manegold, M. Bjørling, W. Cao, J. Gonzalez, J. Granados, N. Hall, S. Idreos, M. Ivanova, R. Johnson, and et al. Repeatability and Workability Evaluation of SIGMOD 2011. *SIGMOD Rec.*, 40(2):45–48, Sept. 2011.

[35] G. C. Bowker and S. L. Star. *Sorting things out: Classification and its consequences*. MIT press, 2000.

[36] N. E. Breslow. Are Statistical Contributions to Medicine Undervalued? *Biometrics*, 59(1):1–8, 2003.

[37] N. E. Breslow. Commentary. *Biostatistics*, 11(3):379–380, July 2010.

[38] A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, S. Kulasekaran, B. Ludäscher, B. D. Mecum, J. Nabrzyski, and et al. Computing environments for reproducibility: Capturing the Whole Tale. *Future Generation Computer Systems*, 94:854–867, 2019.

[39] D. Broockman. Irregularities in LaCour (2014), 2015.

[40] J. B. Buckheit and D. L. Donoho. *WaveLab and Reproducible Research*, pages 55–81. Lecture Notes in Statistics. Springer, 1995.

[41] A. Bugacov, K. Czajkowski, C. Kesselman, A. Kumar, R. E. Schuler, and H. Tangmunarunkit. Experiences with DERIVA: An Asset Management Platform for Accelerating eScience. In *2017 IEEE 13th International Conference on e-Science (e-Science)*, Oct. 2017.

[42] A. Carpen-Amarie, A. Rougier, and F. D. Lübbe. Stepping Stones to Reproducible Research: A Study of Current Practices in Parallel Computing. In L. Lopes, J. Zilinskas, A. Costan, R. G. Cascella, G. Kecskemeti, E. Jeannot, M. Cannataro, L. Ricci, S. Benkner, S. Petit, and et al.Editors, editors, *Euro-Par 2014: Parallel Processing Workshops*, Lecture Notes in Computer Science, pages 499–510. Springer International Publishing, 2014.

[43] A. Casadevall and F. C. Fang. Reproducible Science. *Infection and Immunity*, 78(12):4972–4975, Dec. 2010.

[44] K. K. Cetina. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press, 2009.

[45] J. M. Chambers. Statistical Computing: History and Trends. *The American Statistician*, 34(4):238–243, Nov. 1980.

[46] J. M. Chambers. S as a programming environment for data analysis and graphics. In *Proceedings of the Seventeenth Symposium on the interface of computer sciences and statistics on Computer science and statistics*, pages 211–214. Elsevier North-Holland, Inc., June 1986.

[47] A. C. Chang and P. Li. *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'*. Number 2015-083 in Finance and Economics Discussion Series. Board of Governors of the Federal Reserve System, 2015.

[48] A. C. Chang and P. Li. A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time. *American Economic Review*, 107(5):60–64, May 2017.

[49] D. Chapp, D. Rorabaugh, D. Brown, E. Deelman, K. Vahi, V. Welch, and M. Taufer. Applicability study of the PRIMAD model to LIGO gravitational wave search workflows. *arXiv:1904.05211 [astro-ph]*, Apr. 2019. arXiv: 1904.05211.

[50] D. Chapp, K. Sato, D. H. Ahn, and M. Taufer. Record-and-Replay Techniques for HPC Systems: A Survey. *Supercomputing Frontiers and Innovations*, 5(1):11–30–30, Apr. 2018.

[51] K. Chard, M. D'Arcy, B. Heavner, I. Foster, C. Kesselman, R. Madduri, A. Rodriguez, S. Soiland-Reyes, C. Goble, K. Clark, and et al. I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 319–328, Dec. 2016.

[52] K. Chard, N. Gaffney, M. B. Jones, K. Kowalik, B. Ludäscher, J. Nabrzyski, V. Stodden, I. Taylor, M. J. Turk, and C. Willis. Implementing computational reproducibility in the Whole Tale environment. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, pages 17–22, 2019.

[53] K. Chard, N. Gaffney, M. B. Jones, K. Kowalik, B. Ludäscher, J. Nabrzyski, V. Stodden, I. Taylor, M. J. Turk, and C. Willis. Implementing Computational Reproducibility in the Whole Tale Environment. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS 19, pages 17–22. ACM, 2019. Phoenix, AZ, USA.

[54] B. R. Childers and P. K. Chrysanthis. Artifact Evaluation: FAD or Real News? In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1664–1665, Apr. 2018.

[55] F. Chirigati, R. Capone, R. Rampin, J. Freire, and D. Shasha. A collaborative approach to computational reproducibility. *Information Systems*, 59:95–97, July 2016.

[56] F. Chirigati, R. Rampin, D. Shasha, and J. Freire. ReproZip: Computational Reproducibility With Ease. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD 16, pages 2085–2088. ACM, 2016. event-place: San Francisco, California, USA.

[57] T.-M. Christian, W. G. Jacoby, S. Lafferty-Hess, and T. Carsey. Operationalizing the Replication Standard. *Internal Journal of Digital Curation*, 13(1), 2018.

[58] J. F. Claerbout and M. Karrenbach. *Electronic documents give reproducible research a new meaning*, pages 601–604. Society of Exploration Geophysicists, 1992.

[59] O. S. Collaboration. An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science*, 7(6):657–660, Nov. 2012.

[60] O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), Aug. 2015.

[61] C. Collberg and T. A. Proebsting. Repeatability in Computer Systems Research. *Commun. ACM*, 59(3):62–69, Feb. 2016.

[62] H. Collins. *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press, 1992.

[63] A. E. Commitee. Minutes of the executive committee meetings. *American Economic Review*, 93(2):467–478, May 2003.

[64] D. R. Cox. Biometrika: The First 100 Years. *Biometrika*, 88(1):3–11, 2001.

[65] D. R. Cox and C. Donnelly. Commentary. *Biostatistics*, 11(3):381–382, July 2010.

[66] H. Crowder, R. S. Dembo, and J. M. Mulvey. On Reporting Computational Experiments with Mathematical Software. *ACM Transactions on Mathematical Software (TOMS)*, 5(2):193–203, June 1979.

[67] DA-RT. Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors. Technical report, APSA, 2015.

[68] P. T. Darch. The Core of the Matter: How Do Scientists Judge Trustworthiness of Physical Samples? Under review, 2020.

[69] S. B. Davidson and J. Freire. Provenance and Scientific Workflows: Challenges and Opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD ?08, pages 1345–1350. ACM, 2008. event-place: Vancouver, Canada.

[70] D. De Roure, C. Goble, and R. Stevens. The design and realisation of the Experimentmy Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, May 2009.

[71] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, and et al. Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems, 2005.

[72] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and et al. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*, 46:17–35, May 2015.

[73] T. Delamothe. Quality of websites: kitemarking the west wind: Rating the quality of medical websites may be impossible. *BMJ*, 321(7265):843–844, Oct. 2000.

[74] W. G. Dewald, J. G. Thursby, and R. G. Anderson. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4):587–603, 1986.

[75] K. Diethelm. The Limits of Reproducibility in Numerical Simulation. *Computing in Science Engineering*, 14(1):64–72, Jan. 2012.

[76] P. J. Diggle and S. L. Zeger. Editorial. *Biostatistics*, 10(3):405, July 2009.

[77] P. J. Diggle and S. L. Zeger. Editorial. *Biostatistics*, 11(3):375–375, July 2010.

[78] J. J. Dongarra and E. Grosse. Distribution of mathematical software via electronic mail. *Communications of the ACM*, 30(5):403–407, May 1987.

[79] D. Donoho and V. Stodden. *Reproducible Research in the Mathematical Sciences*, chapter VIII, pages 916–925. Princeton University Press,, 2015.

[80] D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, July 2010.

[81] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden. Reproducible Research in Computational Harmonic Analysis. *Computing in Science Engineering*, 11(1):8–18, Jan. 2009.

[82] C.-L. Douglas. Manuscript Central: Update & Overview. *Editors' Bulletin*, 4(2):79–81, Aug. 2008.

[83] C. Drummond. Reproducible research: a minority opinion. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(1):1–11, Jan. 2018.

[84] E. Duflo and H. Hoynes. Report of the Search Committee to Appoint a Data Editor for the AEA. *AEA Papers and Proceedings*, 108:745, May 2018.

[85] P. N. Edwards. *A vast machine: Computer models, climate data, and the politics of global warming.* MIT Press, 2010.

[86] G. Ellison. Evolving Standards for Academic Publishing: A q-r Theory. *Journal of Political Economy*, 110(5):994–1034, 2002.

[87] N. Eubank. Lessons from a Decade of Replications at the Quarterly Journal of Political Science. *PS: Political Science & Politics*, 49(2):273–276, Apr. 2016.

[88] G. Eysenbach. Thoughts concerning the BMJ editorial 'Kitemarking the west wind' and the WHO dot-health proposal. *Journal of Medical Internet Research*, 2(suppl2):e14, 2000.

[89] D. Fanelli. 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, Mar. 2018.

[90] S. Feigenbaum and D. M. Levy. The market for (ir)reproducible econometrics. *Social Epistemology*, 7(3):215–232, July 1993.

[91] N. Ferro and D. Kelly. SIGIR Initiative to Implement ACM Artifact Review and Badging. *ACM SIGIR Forum*, 52(1):4–10, Aug. 2018.

[92] C. Flaherty. Editorial Malpractice? *Inside Higher Ed*, Apr. 2018.

[93] L. D. Fosdick. Algorithms Policy. *ACM Transactions on Mathematical Software (TOMS)*, 1(1):5–6, Mar. 1975.

[94] J. Freire, P. Bonnet, and D. Shasha. Exploring the coming repositories of reproducible experiments: Challenges and opportunities. *Proceedings of the VLDB Endowment*, pages 1494–1497, Aug. 2011.

[95] J. Freire, N. Fuhr, and A. Rauber. Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041). *Dagstuhl Reports*, 6(1):108–159, 2016.

[96] J. Freire, D. Koop, F. Chirigati, C. T. Silva, D. Koop, F. Chirigati, and C. T. Silva. Reproducibility Using VisTrails, Dec. 2018.

[97] J. Freire and C. T. Silva. Making Computations and Publications Reproducible with VisTrails. *Computing in Science Engineering*, 14(4):18–25, July 2012.

[98] M. Fuentes. Reproducible Research in JASA. *AMSTATNEWS*, July 2016.

[99] G. Fursin and C. Dubach. Community-driven Reviewing and Validation of Publications. In *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering*, TRUST 14, pages 5:1–5:4. ACM, 2014. event-place: Edinburgh, United Kingdom.

[100] G. Fursin, A. Lokhmotov, and E. Plowman. Collective Knowledge: Towards R&D Sustainability. In *Proceedings of the 2016 Conference on Design, Automation & Test in Europe*, DATE 16, pages 864–869. EDA Consortium, 2016. event-place: Dresden, Germany.

[101] G. Fursin, A. Lokhmotov, D. Savenko, and E. Upton. A Collective Knowledge workflow for collaborative research into multi-objective autotuning and machine learning techniques. *CoRR*, abs/1801.08024, 2018.

[102] S. Galiani, P. Gertler, and M. Romero. *Incentives for Replication in Economics.* National Bureau of Economic Research, July 2017.

[103] E. Gallopoulos, E. Houstis, and J. Rice. Computer as thinker/doer: problem-solving environments for computational science. *IEEE Computational Science and Engineering*, 1(2):11–23, 1994.

[104] C. K. Garrett, S. Lien Harrell, and M. A. Heroux. Special Issue on SCC'17 Reproducibility Initiative. *Parallel Computing*, 79:48–49, Nov. 2018.

[105] R. Gentleman and D. T. Lang. Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, 16(1):1–23, May 2007.

[106] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, and et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, Sept. 2004.

[107] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, Aug. 2010.

[108] P. K. Goldberg. Report of the Editor: American Economic Review. *American Economic Review*, 102(3):653–665, May 2012.

[109] P. K. Goldberg. Report of the Editor: American Economic Review. *American Economic Review*, 103(3):701–712, May 2013.

[110] P. K. Goldberg. Report of the Editor: American Economic Review. *American Economic Review*, 107(5):699–712, May 2017.

[111] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341), June 2016.

[112] J. Greenberg, H. C. White, S. Carrier, and R. Scherle. A Metadata Best Practice for a Scientific Data Repository. *Journal of Library Metadata*, 9(3-4):194–212, Nov. 2009.

[113] P. Guo. CDE: A Tool for Creating Portable Experimental Software Packages. *Computing in Science Engineering*, 14(4):32–35, July 2012.

[114] P. J. Guo. CDE: Automatically Package and Reproduce Computational Experiments, Dec. 2018.

[115] R. G. Gutierrez. Stata. *WIREs Computational Statistics*, 2(6):728–733, 2010.

[116] T. Hamasaki and S. Evans. Interview with Professor Geert Molenberghs. *CHANCE*, 30(1):16–23, Jan. 2017.

[117] D. S. Hamermesh. Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue canadienne d?économique*, 40(3):715–733, 2007.

[118] J. Hardwig. The role of trust in knowledge. *The Journal of Philosophy*, 88(12):693–708, 1991.

[119] S. L. Harrell, H. A. Nam, V. G. V. Larrea, K. Keville, and D. Kamalic. Student Cluster Competition: A Multi-Disciplinary Undergraduate HPC Educational Tool. In *Proceedings of the Workshop on Education for High-Performance Computing*, EduHPC ?15. Association for Computing Machinery, 2015. event-place: Austin, Texas.

[120] Y. He and C. H. Q. Ding. Using Accurate Arithmetics to Improve Numerical Reproducibility and Stability in Parallel Applications. *The Journal of Supercomputing*, 18(3):259–277, Mar. 2001.

[121] T. Herndon, M. Ash, and R. Pollin. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics*, 38(2):257–279, 2014.

[122] M. A. Heroux. Editorial: ACM TOMS Replicated Computational Results Initiative. *ACM Trans. Math. Softw.*, 41(3):13:1–13:5, June 2015.

[123] M. A. Heroux. Sustainable & Productive: Improving Incentives for Quality Software. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2016.

[124] M. A. Heroux, L. Barba, M. Parashar, V. Stodden, and M. Taufer. *Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences.* Sandia National Lab, Oct. 2018.

[125] P. S. Herrnson. Replication, Verification, Secondary Analysis, and Data Collection in Political Science. *PS: Political Science and Politics*, 28(3):452–455, 1995.

[126] K. Hinsen. Dealing With Software Collapse. *Computing in Science Engineering*, 21(3):104–108, May 2019.

[127] T. Hoefler and R. Belli. Scientific benchmarking of parallel computing systems: twelve ways to tell the masses when reporting performance results. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC ?15, pages 1–12. Association for Computing Machinery, Nov. 2015.

[128] T. Hopkins. Renovating the Collected Algorithms from ACM. *ACM Transactions on Mathematical Software (TOMS)*, 28(1):59–74, Mar. 2002.

[129] T. Hopkins. The Collected Algorithms of the ACM. *WIREs Computational Statistics*, 1(3):316–324, 2009.

[130] K. Hornik. The Comprehensive R Archive Network. *WIREs Computational Statistics*, 4(4):394–398, 2012.

[131] S. Hunold. A Survey on Reproducibility in Parallel Computing. *arXiv:1511.04217 [cs]*, Nov. 2015. arXiv: 1511.04217.

[132] S. Hunold and J. L. Träff. On the State and Importance of Reproducible Experimental Research in Parallel Computing. *arXiv:1308.3648 [cs]*, Aug. 2013. arXiv: 1308.3648.

[133] R. Ihaka and R. Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[134] J. P. A. Ioannidis. Why Most Published Research Findings Are False, Aug 2005.

[135] W. G. Jacoby, S. Lafferty-Hess, and T.-M. Christian. Should Journals Be Responsible for Reproducibility? *Inside Higher Ed*, July 2017.

[136] D. James, N. Wilkins-Diehr, V. Stodden, D. Colbry, C. Rosales, M. Fahey, J. Shi, R. F. Silva, K. Lee, R. Roskies, and et al. Standing Together for Reproducibility in Large-Scale Computing: Report on reproducibility@XSEDE. *arXiv:1412.5557 [cs]*, Dec. 2014. arXiv: 1412.5557.

[137] M. Jarke and D. Shasha. Information Systems takes a new direction. *Information Systems*, 19(1):1, Jan. 1994.

[138] M. Jarke and D. Shasha. The new Editorial Board of Information Systems. *Information Systems*, 19(2):117–120, Mar. 1994.

[139] B. R. Jasny, G. Chin, L. Chong, and S. Vignieri. *Again, and again, and again?* American Association for the Advancement of Science, 2011.

[140] I. Jimenez, M. Sevilla, N. Watkins, C. Maltzahn, J. Lofstead, K. Mohror, A. Arpaci-Dusseau, and R. Arpaci-Dusseau. The Popper Convention: Making Reproducible Systems Evaluation Practical. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1561–1570, May 2017.

[141] N. Keiding. Reproducible research and the substantive context. *Biostatistics*, 11(3):376–378, July 2010.

[142] N. Keiding. Reproducible research and the substantive context: response to comments. *Biostatistics*, 11(3):395–396, July 2010.

[143] E. M. Key. How Are We Doing? Data Access and Replication in Political Science. *PS: Political Science & Politics*, 49(2):268–272, Apr. 2016.

[144] G. King. A Revised Proposal, Proposal. *PS: Political Science and Politics*, 28(3):494–499, 1995.

[145] G. King. Replication, Replication. *PS: Political Science & Politics*, 28(3):444–452, Sept. 1995.

[146] G. King. The Future of Replication. *International Studies Perspectives*, 4:443–499, 2003.

[147] G. King. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods and Research*, 36:173–199, 2007.

[148] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, et al. Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90, 2016.

[149] K. D. Knorr-Cetina. Scientific Communities or Transepistemic Arenas of Research? A Critique of Quasi-Economic Models of Science. *Social Studies of Science*, 12(1):101–130, Feb. 1982.

[150] D. E. Knuth. Literate programming. *The Computer Journal*, 27(2):97–111, 1984.

[151] R. Koenker and A. Zeileis. Reproducible Econometric Research. A Critical Review of the State of the Art., 2007.

[152] H. Koers and R. Capone. New article type verifies experimental reproducibility, Apr. 2016.

[153] C. Kooperberg. StatLib: An Archive for Statistical Software, Datasets, and Information. *The American Statistician*, 51(1):98, Feb. 1997.

[154] J. Kovacevic. How to Encourage and Publish Reproducible Research. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ?07*, volume 4, pages IV–1273–IV–1276, Apr. 2007.

[155] M. Krafczyk, A. Shi, A. Bhaskar, D. Marinov, and V. Stodden. Scientific Tests and Continuous Integration Strategies to Enhance Reproducibility in the Scientific Software Context. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS 19, pages 23–28. Association for Computing Machinery, June 2019.

[156] S. Krishnamurthi. Artifact Evaluation for Software Conferences. *SIGSOFT Softw. Eng. Notes*, 38(3):7–10, May 2013.

[157] S. Krishnamurthi and J. Vitek. The Real Software Crisis: Repeatability As a Core Value. *Commun. ACM*, 58(3):34–36, Feb. 2015.

[158] G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5), 2017.

[159] S. Kvale. *InterViews: An Introduction to Qualitative Research Interviewing.* SAGE Publications, Inc, 1996.

[160] C. Laine, S. N. Goodman, M. E. Griswold, and H. C. Sox. Reproducible Research: Moving toward Research the Public Can Really Trust. *Annals of Internal Medicine*, 146(6):450, Mar. 2007.

[161] J. T. Leek and R. D. Peng. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646, Feb. 2015.

[162] S. Leigh Star. This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, and Human Values*, 35(5):601–617, Sept. 2010.

[163] F. Leisch. Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In W. Härdle and B. Rönz, editors, *Compstat*, pages 575–580. Physica-Verlag HD, 2002.

[164] B. Lerner and E. Boose. RDataTracker: collecting provenance in an interactive scripting environment. In *6th {USENIX} Workshop on the Theory and Practice of Provenance (TaPP 2014)*, 2014.

[165] R. J. LeVeque. Wave propagation software, computational science, and reproducible research. In *Proc. Int. Congr. of Mathematician*, 2006.

[166] R. J. LeVeque, I. M. Mitchell, and V. Stodden. Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science Engineering*, 14(4):13–17, July 2012.

[167] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the Kepler system: Research Articles. *Concurrency and Computation: Practice & Experience*, 18(10):1039–1065, Aug. 2006.

[168] A. Lupia and C. Elman. Openness in Political Science: Data Access and Research Transparency: Introduction. *PS: Political Science & Politics*, 47(1):19–42, Jan. 2014.

[169] J. Lyle. OpenICPSR. *Bulletin of the Association for Information Science and Technology*, 40(5):55–56, 2014.

[170] J. MacKinnon. Guidelines for Users Journal of Applied Econometrics FTP Archive Site. *Journal of Applied Econometrics*, 9(2):229–230, 1994.

[171] D. Madigan and R. Wasserstein. *Statistics and science: a report of the London workshop on the future of the statistical sciences.* World of Statistics, 2014.

[172] A. Magi, M. Benelli, G. Marseglia, G. Nannetti, M. R. Scordo, and F. Torricelli. A shifting level model algorithm that identifies aberrations in array-CGH data. *Biostatistics*, 11(2):265–280, Apr. 2010.

[173] S. Manegold, I. Manolescu, L. Afanasiev, J. Feng, G. Gou, M. Hadjieleftheriou, S. Harizopoulos, P. Kalnis, K. Karanasos, D. Laurent, and et al. Repeatability & Workability Evaluation of SIGMOD 2009. *SIGMOD Rec.*, 38(3):40–43, Dec. 2010.

[174] I. Manolescu, L. Afanasiev, A. Arion, J. Dittrich, S. Manegold, N. Polyzotis, K. Schnaitter, P. Senellart, S. Zoupanos, and D. Shasha. The Repeatability Experiment of SIGMOD 2008. *SIGMOD Rec.*, 37(1):39–45, Mar. 2008.

[175] B. Marwick. rrtools. `https://github.com/benmarwick/rrtools`, 2020.

[176] B. Marwick, C. Boettiger, and L. Mullen. Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician*, 72(1):80–88, Jan. 2018.

[177] B. D. McCullough. Assessing the Reliability of Statistical Software: Part I. *The American Statistician*, 52(4):358–366, Nov. 1998.

[178] B. D. McCullough. Assessing the Reliability of Statistical Software: Part II. *The American Statistician*, 53(2):149–159, May 1999.

[179] B. D. McCullough, K. A. McGeary, and T. D. Harrison. Lessons from the JMCB Archive. *Journal of Money, Credit, and Banking*, 38(4):1093–1107, 2006.

[180] B. D. McCullough, K. A. McGeary, and T. D. Harrison. Do Economics Journal Archives Promote Replicable Research? *The Canadian Journal of Economics / Revue canadienne d?Economique*, 41(4):1406–1420, 2008.

[181] B. D. McCullough and H. D. Vinod. The Numerical Reliability of Econometric Software. *Journal of Economic Literature*, 37(2):633–665, 1999.

[182] B. D. McCullough and H. D. Vinod. Verifying the Solution from a Nonlinear Solver: A Case Study. *The American Economic Review*, 93(3):873–892, 2003.

[183] T. McPhillips, T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, R. K. Bocinsky, Y. Cao, J. Cheney, F. Chirigati, S. Dey, and et al. YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. *International Journal of Digital Curation*, 10(1):298–313, Feb. 2015.

[184] K. J. Meier. Replication: A View from the Streets. *PS: Political Science and Politics*, 28(3):456–459, 1995.

[185] H. Meng, R. Kommineni, Q. Pham, R. Gardner, T. Malik, and D. Thain. An invariant framework for conducting reproducible computational science. *Journal of Computational Science*, 9:137–142, July 2015.

[186] G. W. Miller. Reproducibility Revisited: Reflections of an Editor. *Toxicological Sciences*, 169(2):315–316, June 2019.

[187] P. Mirowski and S. Sklivas. Why econometricians don?t replicate (although they do reproduce). *Review of Political Economy*, 3(2):146–163, Apr. 1991.

[188] R. A. Moffit. Report of the Editor. *American Economic Review*, 96(2):497–509, May 2006.

[189] R. A. Moffitt. Report of the Editor: American Economic Review (with Appendix by Philip J. Glandon). *American Economic Review*, 101(3):684–693, May 2011.

[190] P. Montesinos, L. Ceze, and J. Torrellas. DeLorean: Recording and Deterministically Replaying Shared-Memory Multiprocessor Execution Ef?ciently. *ACM SIGARCH Computer Architecture News*, 36(3):289–300, June 2008.

[191] L. Murta, V. Braganholo, F. Chirigati, D. Koop, and J. Freire. noWorkflow: Capturing and Analyzing Provenance of Scripts. In B. Ludäscher and B. Plale, editors, *Provenance and Annotation of Data and Processes*, Lecture Notes in Computer Science, pages 71–83. Springer International Publishing, 2015.

[192] J. A. Nelder and B. E. Cooper. Epilogue. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 16(2):149–151, 1967.

[193] N. H. Nie, D. H. Bent, and C. H. Hull. *SPSS: Statistical package for the social sciences*, volume 227. McGraw-Hill New York, 1975.

[194] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, and et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[195] B. A. Nosek, J. R. Spies, and M. Motyl. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, 7(6):615–631, Nov. 2012.

[196] D. Nüst, C. Boettiger, and B. Marwick. How to Read a Research Compendium. *arXiv:1806.09525 [cs]*, June 2018. arXiv: 1806.09525.

[197] L. Oliveira, D. Wilkinson, D. Mossé, and B. Childers. Supporting Long-term Reproducible Software Execution. In *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS'18, pages 6:1–6:6. ACM, 2018. event-place: Tempe, AZ, USA.

[198] Y. L. S. R. on Data and C. Sharing. Reproducible Research. *Computing in Science Engineering*, 12(5):8–13, Sept. 2010.

[199] C. on Reproducibility and R. in Science. *Reproducibility and Replicability in Science.* National Academies Press, 2019.

[200] T. W. P. on Statistical Computing. The Construction and Description of Algorithms. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 17(2):175–179, 1968.

[201] T. W. P. on Statistical Computing. The Construction and Description of Algorithms. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):366–373, 1975.

[202] T. W. P. on Statistical Computing. The Construction and Description of Algorithms. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(3):311–318, 1979.

[203] A. S. Palus. Conservative political beliefs not linked to psychotic traits, as study claimed, June 2016.

[204] M. Parashar. Editor?s Note: IEEE Transactions on Parallel and Distributed Systems (TPDS) Reproducibility Initiative, June 2019. *IEEE Transactions on Parallel and Distributed Systems*, 30(8):1690–1690, Aug. 2019.

[205] W. S. Parker. Does matter really matter? Computer simulations, experiments, and materiality. *Synthese*, 169(3):483–496, Aug. 2009.

[206] T. Pasquier, M. K. Lau, X. Han, E. Fong, B. S. Lerner, E. Boose, M. Crosas, A. M. Ellison, and M. Seltzer. Sharing and Preserving Computational Analyses for Posterity with encapsulator. *arXiv:1803.05808 [cs]*, May 2018. arXiv: 1803.05808.

[207] D. Patel. Research data management: a conceptual framework. *Library Review*, 65(4/5):226–241, Jan. 2016.

[208] M. Q. Patton. Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34(5 Pt 2):1189–1208, Dec. 1999.

[209] R. Peng. The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32, 2015.

[210] R. Peng, L. Welty, and A. McDermott. The National Morbidity, Mortality, and Air Pollution Study Database in R. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, June 2004.

[211] R. D. Peng. Reproducible research and Biostatistics. *Biostatistics*, 10(3):405–408, July 2009.

[212] R. D. Peng. Discussion of Keiding. *Biostatistics*, 11(3):393–394, July 2010.

[213] R. D. Peng. Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227, Dec. 2011.

[214] R. D. Peng, F. Dominici, R. Pastor-Barriuso, S. L. Zeger, and J. M. Samet. Seasonal Analyses of Air Pollution and Mortality in 100 US Cities. *American Journal of Epidemiology*, 161(6):585–594, Mar. 2005.

[215] R. D. Peng, F. Dominici, and S. L. Zeger. Reproducible Epidemiologic Research. *American Journal of Epidemiology*, 163(9):783–789, May 2006.

[216] H. Pesaran. Introducing a replication section. *Journal of Applied Econometrics*, 18(1):111–111, 2003.

[217] J. Pevehouse. Editor?s Note. *International Organization*, 66(3):359–362, 2012.

[218] Q. Pham, T. Malik, and I. Foster. Using provenance for repeatability. In *Presented as part of the 5th {USENIX} Workshop on the Theory and Practice of Provenance*, 2013.

[219] H. E. Plesser. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11, 2018.

[220] K. Popper. *The Logic of Scientific Discovery*. Routledge, 2002.

[221] L. Pouchard, S. Baldwin, T. Elsethagen, S. Jha, B. Raju, E. Stephan, L. Tang, and K. K. Van Dam. Computational reproducibility of scientific workflows at extreme scales. *The International Journal of High Performance Computing Applications*, page 1094342019839124, Apr. 2019.

[222] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, Sept. 2011.

[223] Project TIER. Project TIER. https://www.projecttier.org/.

[224] A. Purcell. Tool developed at CERN makes software citation easier, 2014.

[225] H. Radder. *In and about the world: Philosophical studies of science and technology*. SUNY Press, 1996.

[226] H. Radder. Which Scientific Knowledge is a Common Good? *Social Epistemology*, 31(5):431–450, Sept. 2017.

[227] D. B. Resnik and A. E. Shamoo. Reproducibility and Research Integrity. *Accountability in research*, 24(2):116–123, 2017.

[228] J. Rice and R. Boisvert. From scientific software libraries to problem-solving environments. *IEEE Computational Science and Engineering*, 3(3):44–53, 1996.

[229] J. R. Rice. Purpose and Scope. *ACM Transactions on Mathematical Software (TOMS)*, 1(1):1–3, Mar. 1975.

[230] J. R. Rice and R. F. Boisvert. *Scalable Software Libraries and Problem Solving Environments*, pages 33–43. The Springer International Series in Engineering and Computer Science. Springer US, 2000.

[231] D. S. H. Rosenthal. Emulation & Virtualization as Preservation Strategies, 2015.

[232] A. J. Rossini. Literate Statistical Practice. In *Proceedings of DSC*, 2001.

[233] A. Rowhani-Farid and A. G. Barnett. Badges for sharing data and code at Biostatistics: an observational study. *F1000Research*, 7, Mar. 2018.

[234] N. Salkind. Journal of the American Statistical Association, 2007.

[235] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology*, 9(10):e1003285, Oct. 2013.

[236] M. Schreier. *Qualitative content analysis in practice.* London; Thousand Oaks, Calif.: Sage Publications Ltd, 2012., 2012.

[237] R. E. Schuler, C. Kesselman, and K. Czajkowski. Accelerating data-driven discovery with scientific asset management. In *2016 IEEE 12th International Conference on e-Science (e-Science)*, pages 31–40, Oct. 2016.

[238] M. Schwab, N. Karrenbach, and J. Claerbout. Making scientific computations reproducible. *Computing in Science Engineering*, 2(6):61–67, Nov. 2000.

[239] J. Sedransk, P. Switzer, and J. M. Tanur. Editors' Report for 1987. *Journal of the American Statistical Association*, 83(402):287–288, 1988.

[240] P. Sefton, E. O. Carragain, S. Soiland-Reyes, O. Corcho, D. Garijo, R. Palma, F. Coppens, C. Goble, J. M. Fernandez, K. Chard, and et al. RO-Crate Metadata Specification 1.0, Nov. 2019.

[241] D. Shasha. Editorial. *Information Systems*, 29(3):205, May 2004.

[242] S. Soiland-Reyes, M. Gamble, and R. Haines. *Research Object Bundle 1.0, researchobject.org Recommendation.* Zenodo, 2014.

[243] S. Soltesz, H. Pötzl, M. E. Fiuczynski, A. Bavier, and L. Peterson. Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors. In *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, EuroSys ?07, pages 275–287. Association for Computing Machinery, Mar. 2007.

[244] D. Spiegelhalter. Trust in numbers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):948–965, 2017.

[245] A. T. Staff. Algorithms Distribution Service. *ACM Transactions on Mathematical Software (TOMS)*, 1(1):4, Mar. 1975.

[246] S. L. Star and J. R. Griesemer. 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, 1989.

[247] S. L. Star and K. Ruhleder. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, Mar. 1996.

[248] C. A. Stewart, T. M. Cockerill, I. Foster, D. Hancock, N. Merchant, E. Skidmore, D. Stanzione, J. Taylor, S. Tuecke, G. Turner, et al. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, pages 1–8, 2015.

[249] J. Stimson, C. Franklin, W. R. Mebane, P. A. Schrodt, and B. D. Wood. Statement on Statistcal Reporting, Archiving and Replication: Norms for Publicaiton. *The Political Methodologist*, 6(1):18–19, 1994.

[250] V. Stodden. The Legal Framework for Reproducible Scientific Research: Licensing and Copyright. *Computing in Science Engineering*, 11(1):35–40, Jan. 2009.

[251] V. Stodden. Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science. *Computing in Science Engineering*, 12(5):8–12, 2010.

[252] V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics, 2013.

[253] V. Stodden, P. Guo, and Z. Ma. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLOS ONE*, 8(6):e67111, June 2013.

[254] V. Stodden and M. S. Krafczyk. Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context. In *1st Workshop on Reproducible, Customizable and Portable Workflows for HPC*, 2018.

[255] V. Stodden, M. S. Krafczyk, and A. Bhaskar. Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility. In *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS'18, pages 3:1–3:5. ACM, 2018. event-place: Tempe, AZ, USA.

[256] V. Stodden, S. Miguez, and J. Seiler. ResearchCompendia.org: Cyberinfrastructure for Reproducibility and Collaboration in Computational Science. *Computing in Science and Engineering*, 17(1):12–19, 2015.

[257] V. Stodden, J. Seiler, and Z. Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, Mar. 2018.

[258] M. Taufer. SC16 Explores Reproducibility for Advanced Computing Through Student Cluster Competition, Mar. 2016.

[259] M. Taufer, O. Padron, P. Saponaro, and S. Patel. Improving numerical reproducibility and stability in large-scale numerical simulations on GPUs. In *2010 IEEE International Symposium on Parallel Distributed Processing (IPDPS)*, pages 1–9, Apr. 2010.

[260] J. Tennant, J. Dugan, D. Graziotin, D. Jacques, F. Waldner, D. Mietchen, Y. Elkhatib, L. B. Collister, C. Pikas, T. Crick, and et al. A multi-disciplinary perspective on emergent and future innovations in peer review [version 3; peer review: 2 approved]. *F1000Research*, 6(1151), 2017.

[261] C. Tenopir, K. Levine, S. Allard, L. Christian, R. Volentine, R. Boehm, F. Nichols, D. Nicholas, H. R. Jamali, E. Herman, and et al. Trustworthiness and authority of scholarly information in a digital age: Results of an international questionnaire. *Journal of the Association for Information Science and Technology*, 67(10):2344–2361, 2016.

[262] D. H. T. That, G. Fils, Z. Yuan, and T. Malik. Sciunits: Reusable Research Objects. *CoRR*, abs/1707.05731, 2017.

[263] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering*, 16(5):62–74, Sept. 2014.

[264] P. Vandewalle. Code Sharing Is Associated with Research Impact in Image Processing. *Computing in Science Engineering*, 14(4):42–47, July 2012.

[265] P. Vandewalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26(3):37–47, May 2009.

[266] L. Vilhuber. *Reproducibility and Replicability in Economics.* National Academies of Science, Engineering, and Medicine, 2018.

[267] L. Vilhuber. Report by the AEA Data Editor. *AEA Papers and Proceedings*, 109:718–729, May 2019.

[268] L. Vilhuber, J. Turrito, and K. Welch. Report by the AEA Data Editor. *AEA Papers and Proceedings*, 110, May 2020.

[269] J. Vitek and T. Kalibera. Repeatability, reproducibility and rigor in systems research. In *2011 Proceedings of the Ninth ACM International Conference on Embedded Software (EMSOFT)*, pages 33–38, Oct. 2011.

[270] J. Ware. The Kitemark - its history and benefits. *International Journal of Technology Management*, 1(3-4):491–497, Jan. 1986.

[271] R. L. Wasserstein. Addressing Reproducible Research: How the Statistical Community Can Help (and Has Been Helping), Nov. 2017.

[272] R. L. Wasserstein and N. A. Lazar. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, Apr. 2016.

[273] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, and I. Foster. Swift: A language for distributed parallel scripting. *Parallel Computing*, 37(9):633–652, Sept. 2011.

[274] T. Wilholt. Epistemic Trust in Science. *The British Journal for the Philosophy of Science*, 64(2):233–253, June 2013.

[275] G. R. Williams, G. P. Behm, T. Nguyen, A. Esparza, V. G. Haka, A. Ramos, B. Wright, J. C. Otto, C. P. Paolini, and M. P. Thomas. SC16 student cluster competition challenge: Investigating the reproducibility of results for the ParConnect application. *Parallel Computing*, 70:27–34, Dec. 2017.

[276] R. Wilson. Note from the Editor. *American Journal of Political Science*, 56(3):519–519, 2012.

[277] E. Winsberg. *Science in the age of computer simulation.* University of Chicago Press, 2010.

[278] Y. Xie. knitr: A Comprehensive Tool for Reproducible Research in R, Dec. 2018.

[279] R. K. Yin. *Case study research and applications: Design and methods.* Sage publications, 2017.

[280] S. L. Zeger. Editorial: Knowledge from Information. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 170(3):513–516, 2007.

Appendix A

NASEM Recommendations

This appendix provides the text of the four recommendations (4-1, 6-3, 6-4, 6-5) from the National Academics of Sciences, Engineering, and Medicine's Committee (NASEM) on Reproducibility and Replicability in Science [199] that are the focus of the study presented here.

RECOMMENDATION 4-1

> To help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment:
>
> 1. the input data used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;
> 2. a detailed description of the study methods (ideally in executable form) together with its computational steps and associated parameters; and
> 3. information about the computational environment where the study was originally executed, such as operating system, hardware architecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

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:

1. Develop a set of criteria for trusted open repositories to be used by the scientific community for objects of the scholarly record.
2. Seek to harmonize with other funding agencies the repository criteria and data-management plans for scholarly objects.
3. 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.
4. Consider extending NSF's current data-management plan to include other digital artifacts, such as software.
5. Work with communities reliant on non-public data or code to develop alternative mechanisms for demonstrating reproducibility

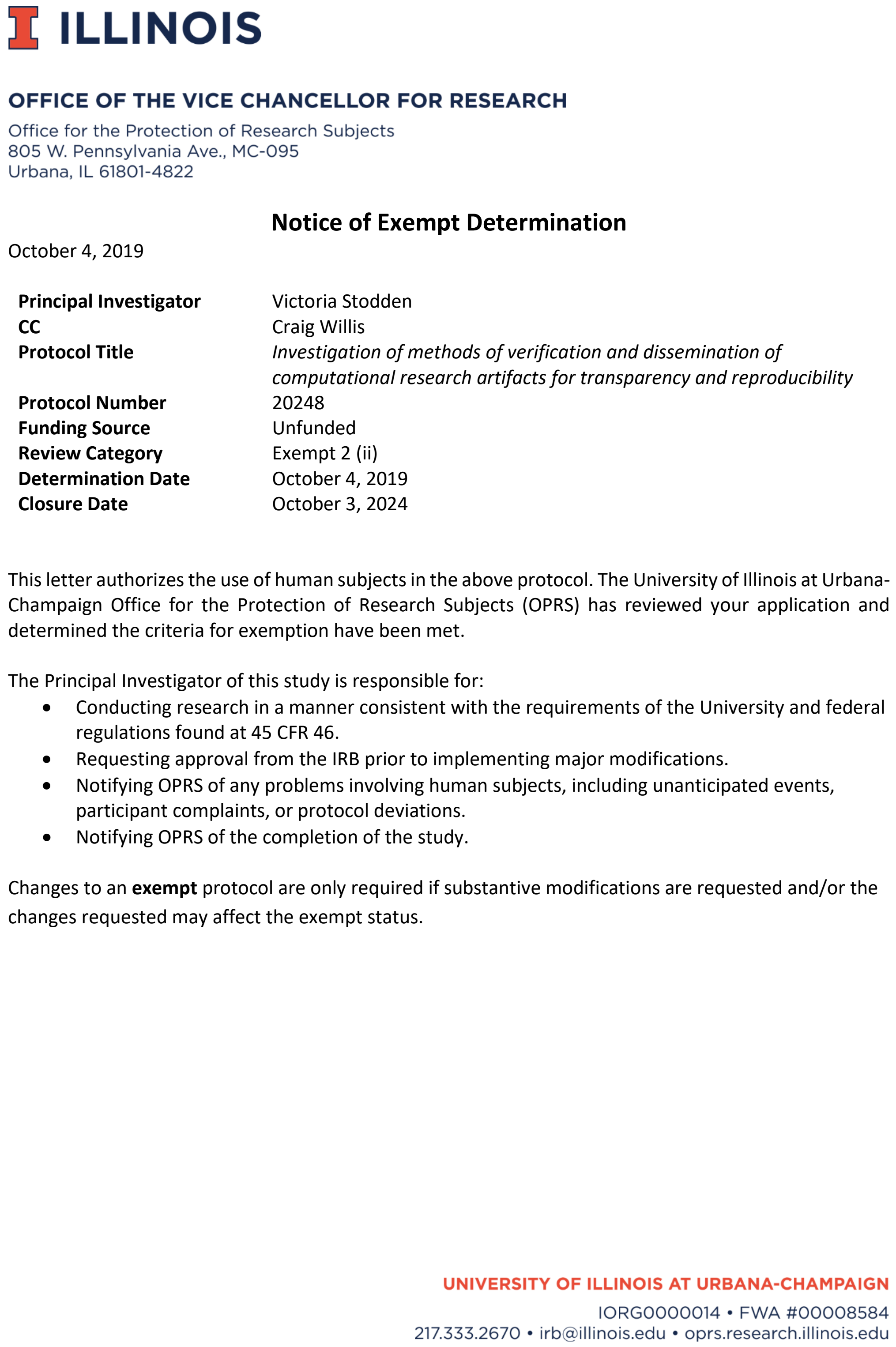
Through these repository criteria, NSF would enable discoverability and standards for digital scholarly objects and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

Appendix B

IRB Materials

This appendix includes the Institutional Review Board (IRB) materials for this study including exempt determination, recruitment materials, informed consent, and interview protocol.

B.1 Exempt Determination



ILLINOIS

OFFICE OF THE VICE CHANCELLOR FOR RESEARCH

Office for the Protection of Research Subjects
805 W. Pennsylvania Ave., MC-095
Urbana, IL 61801-4822

Notice of Exempt Determination

October 4, 2019

Principal Investigator	Victoria Stodden
CC	Craig Willis
Protocol Title	*Investigation of methods of verification and dissemination of computational research artifacts for transparency and reproducibility*
Protocol Number	20248
Funding Source	Unfunded
Review Category	Exempt 2 (ii)
Determination Date	October 4, 2019
Closure Date	October 3, 2024

This letter authorizes the use of human subjects in the above protocol. The University of Illinois at Urbana-Champaign Office for the Protection of Research Subjects (OPRS) has reviewed your application and determined the criteria for exemption have been met.

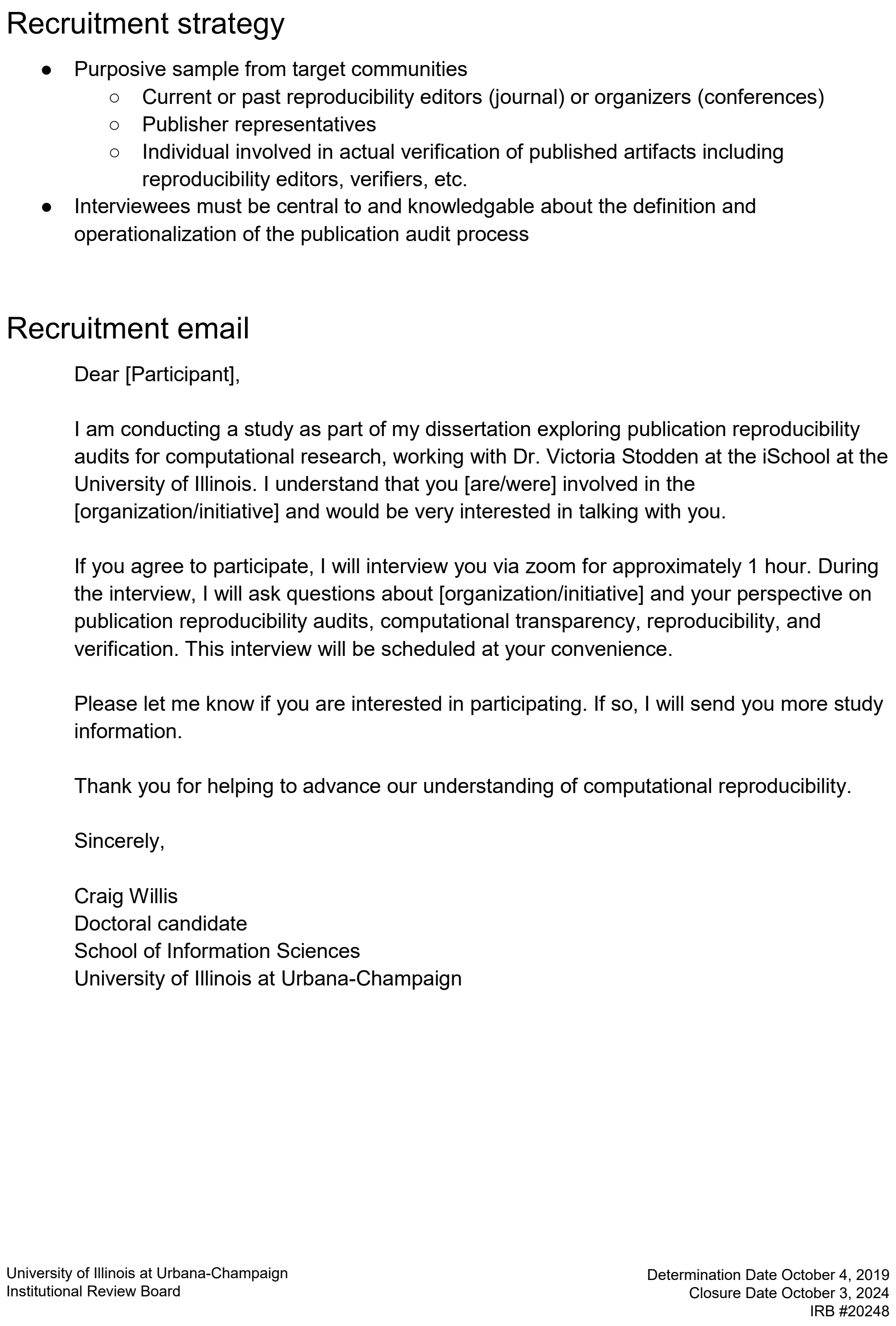
The Principal Investigator of this study is responsible for:

- Conducting research in a manner consistent with the requirements of the University and federal regulations found at 45 CFR 46.
- Requesting approval from the IRB prior to implementing major modifications.
- Notifying OPRS of any problems involving human subjects, including unanticipated events, participant complaints, or protocol deviations.
- Notifying OPRS of the completion of the study.

Changes to an **exempt** protocol are only required if substantive modifications are requested and/or the changes requested may affect the exempt status.

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
IORG0000014 • FWA #00008584
217.333.2670 • irb@illinois.edu • oprs.research.illinois.edu

B.2 Recruitment Materials



Recruitment strategy

- Purposive sample from target communities
  - Current or past reproducibility editors (journal) or organizers (conferences)
  - Publisher representatives
  - Individual involved in actual verification of published artifacts including reproducibility editors, verifiers, etc.
- Interviewees must be central to and knowledgable about the definition and operationalization of the publication audit process

Recruitment email

Dear [Participant],

I am conducting a study as part of my dissertation exploring publication reproducibility audits for computational research, working with Dr. Victoria Stodden at the iSchool at the University of Illinois. I understand that you [are/were] involved in the [organization/initiative] and would be very interested in talking with you.

If you agree to participate, I will interview you via zoom for approximately 1 hour. During the interview, I will ask questions about [organization/initiative] and your perspective on publication reproducibility audits, computational transparency, reproducibility, and verification. This interview will be scheduled at your convenience.

Please let me know if you are interested in participating. If so, I will send you more study information.

Thank you for helping to advance our understanding of computational reproducibility.

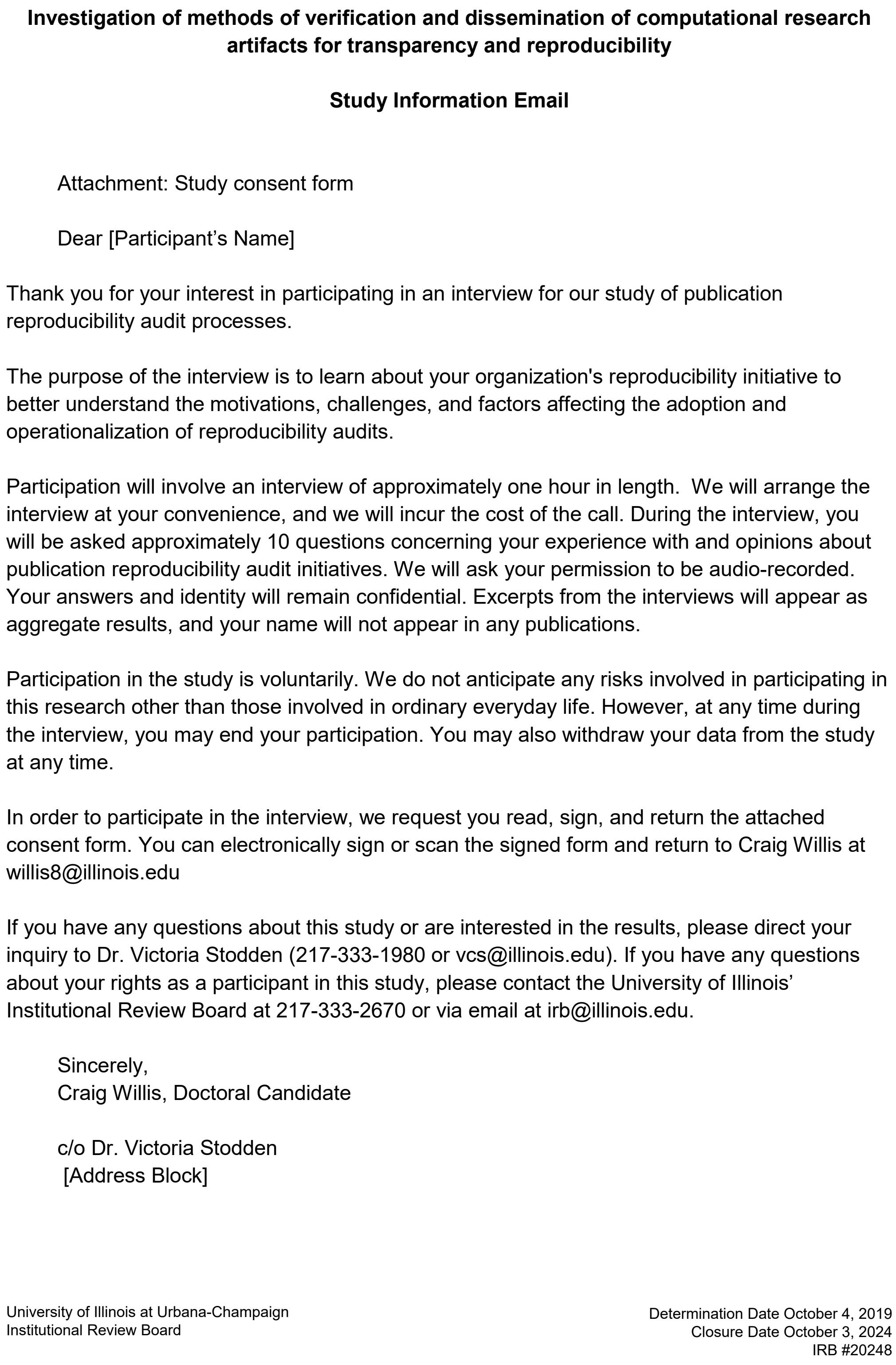
Sincerely,

Craig Willis
Doctoral candidate
School of Information Sciences
University of Illinois at Urbana-Champaign

University of Illinois at Urbana-Champaign
Institutional Review Board

Determination Date October 4, 2019
Closure Date October 3, 2024
IRB #20248

B.3 Informed Consent



Investigation of methods of verification and dissemination of computational research artifacts for transparency and reproducibility

Study Information Email

Attachment: Study consent form

Dear [Participant's Name]

Thank you for your interest in participating in an interview for our study of publication reproducibility audit processes.

The purpose of the interview is to learn about your organization's reproducibility initiative to better understand the motivations, challenges, and factors affecting the adoption and operationalization of reproducibility audits.

Participation will involve an interview of approximately one hour in length. We will arrange the interview at your convenience, and we will incur the cost of the call. During the interview, you will be asked approximately 10 questions concerning your experience with and opinions about publication reproducibility audit initiatives. We will ask your permission to be audio-recorded. Your answers and identity will remain confidential. Excerpts from the interviews will appear as aggregate results, and your name will not appear in any publications.

Participation in the study is voluntarily. We do not anticipate any risks involved in participating in this research other than those involved in ordinary everyday life. However, at any time during the interview, you may end your participation. You may also withdraw your data from the study at any time.

In order to participate in the interview, we request you read, sign, and return the attached consent form. You can electronically sign or scan the signed form and return to Craig Willis at willis8@illinois.edu

If you have any questions about this study or are interested in the results, please direct your inquiry to Dr. Victoria Stodden (217-333-1980 or vcs@illinois.edu). If you have any questions about your rights as a participant in this study, please contact the University of Illinois' Institutional Review Board at 217-333-2670 or via email at irb@illinois.edu.

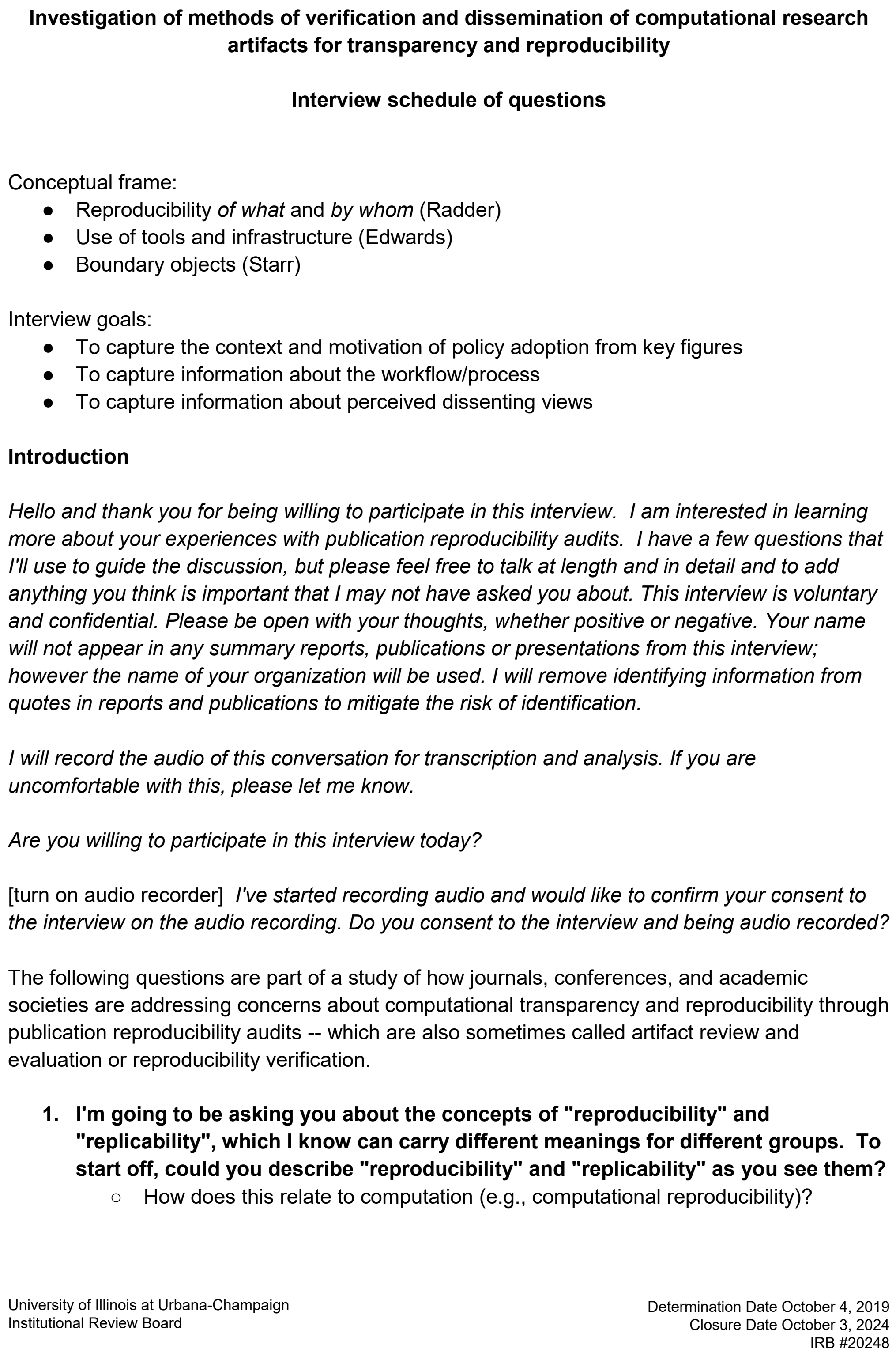
Sincerely,
Craig Willis, Doctoral Candidate

c/o Dr. Victoria Stodden
[Address Block]

University of Illinois at Urbana-Champaign
Institutional Review Board

Determination Date October 4, 2019
Closure Date October 3, 2024
IRB #20248

B.4 Interview Protocol



Investigation of methods of verification and dissemination of computational research artifacts for transparency and reproducibility

Interview schedule of questions

Conceptual frame:

- Reproducibility *of what* and *by whom* (Radder)
- Use of tools and infrastructure (Edwards)
- Boundary objects (Starr)

Interview goals:

- To capture the context and motivation of policy adoption from key figures
- To capture information about the workflow/process
- To capture information about perceived dissenting views

Introduction

Hello and thank you for being willing to participate in this interview. I am interested in learning more about your experiences with publication reproducibility audits. I have a few questions that I'll use to guide the discussion, but please feel free to talk at length and in detail and to add anything you think is important that I may not have asked you about. This interview is voluntary and confidential. Please be open with your thoughts, whether positive or negative. Your name will not appear in any summary reports, publications or presentations from this interview; however the name of your organization will be used. I will remove identifying information from quotes in reports and publications to mitigate the risk of identification.

I will record the audio of this conversation for transcription and analysis. If you are uncomfortable with this, please let me know.

Are you willing to participate in this interview today?

[turn on audio recorder] *I've started recording audio and would like to confirm your consent to the interview on the audio recording. Do you consent to the interview and being audio recorded?*

The following questions are part of a study of how journals, conferences, and academic societies are addressing concerns about computational transparency and reproducibility through publication reproducibility audits -- which are also sometimes called artifact review and evaluation or reproducibility verification.

1. **I'm going to be asking you about the concepts of "reproducibility" and "replicability", which I know can carry different meanings for different groups. To start off, could you describe "reproducibility" and "replicability" as you see them?**
    - How does this relate to computation (e.g., computational reproducibility)?

University of Illinois at Urbana-Champaign
Institutional Review Board

Determination Date October 4, 2019
Closure Date October 3, 2024
IRB #20248

    - Please describe "transparency" in this same context and how it relates to reproducibility

2. **Tell me about the [organization reproducibility initiative].**
    - What does the process look like today?
    - Is it still actively used?
    - Where do you believe that the [organization reproducibility initiative] fits with the definitions you provided? Is it concerned with reproducibility, replicability, transparency, or something else?
3. **I'm interested in the history of this initiative**
    - How did it start?
    - What problem is the community trying to solve?
    - Do you believe that it solves the problem as intended? If so, how?
    - Is re-use or extensibility a concern? Please explain.
    - Who benefits from this process and how?
4. **I'd like to know about your research community's reaction to this initiative**
    - How have researchers responded?
    - What are their key concerns, if any?
    - What are your key concerns, if any?
5. **I'd like to talk about the mechanics of the process**
    - Is the process voluntary or mandatory? Why?
    - Can you briefly describe the workflow for a paper?
    - Has the process changed over time?
    - How do you decide what should be reproduced in a paper? What is a "result" or "key finding"?
    - Is there a documented workflow or guidelines for reviewers? Is it versioned? Is it published?
    - Do you produce reports for editors or authors? How are they handled?
    - Are artifacts published to archival repositories? Is this important?
    - How do you measure or envision measuring the impact of this initiative?
6. **I'd like to talk about required skills of both researchers and reviewers**
    - What kinds of new skills does the researcher need to be successful through this process?
    - What makes a good verifier/reviewer?
        - What are the skills and experiences that a verifier needs to do this work?
        - Do reviewers need to understand the theory or methods?
7. **I'd like to get your perspective on the value of reproducibility audits**
    - What are the benefits of reproducibility audits?
    - How have you benefited from being involved in this initiative?
    - What are the challenges?
    - What should other organizations learn from what you've done?

Thank you for your time and for participating in our study. For the purpose of describing the participant pool in reports, I have a few demographic questions.

8. **Demographic questions**
    - What is the title of your current position?
    - How long have you been in this position?
    - What is your degree and field?

Appendix C

Codebooks

This appendix includes the codebooks used for the qualitative analysis of interview transcripts and documentary evidence.

C.1 Codes used for 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 themself
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 C.1: High-level code groups used for coding of interview transcripts

Code	**Definition**
Badges	Discussionof badges and badging or related metadata/statements/assertions
Benefits: Archive	Discussion of benefits to the archive or repository
Benefits: Author	Discussion of benefits to the author/researcher
Benefits: Community	Discussion of benefits to the community/discipline
Benefits: Curator	Discussion of benefits to the curator or archivist
Benefits: Funders	Discussion of benefits to funders
Benefits: Journal	Discussion of benefits to the journal

Code	**Definition**
Benefits: Reviewers	Discussion of benefits to reviewers
Benefits: Self	Discussion of benefits to self
Benefits: Students	Discussion of benefits to students
Benefits: Verifiers	Discussion of benefits to verifiers
Challenges: Awareness: Archive	Discussion of challenges related to archive staff awareness or understanding of policy.
Challenges: Awareness: Author	Discussion of challenges related to author awareness or understanding of policy.
Challenges: Awareness: Community	Discussion of challenges related to community awareness or understanding of policy.
Challenges: Awareness: Editors	Discussion of challenges related to editorial team awareness/understanding
Challenges: Burden: Author	Discussion of challenges/concerns related to burden placed on authors
Challenges: Burden: Curator	Discussion of challenges/concerns related to burden placed on curators
Challenges: Burden: Editor	Discussion of challenges/concerns related to burden placed on editors
Challenges: Burden: Reviewer	Discussion of challenges/concerns related to burden placed on reviewers
Challenges: BuyIn	Discussion of challenges related to stakeholder buy-in to initiative
Challenges: Cost	Discussion of challenges related to cost
Challenges: EditorialChange	Discussion of challenges related to actual editorial leadership changes
Challenges: Efficiency	Discussion of challenges related to review workflow efficiency
Challenges: Environment	Discussion of challenges related to computational envrionment
Challenges: Expertise	Discussion of challanges related to expertise of stakeholders
Challenges: Infrastructure: Archive	Discussion of challanges related to archiving and archiving infrastructure
Challenges: Infrastructure: Editorial	Discussion of challenges related to lack of administrative infrastructure
Challenges: Infrastructure: Licenses	Discussion of challenges related to access to licenses
Challenges: Infrastructure: Publishing Platform	Discussion of challenges related to publishing platform limitations
Challenges: Infrastructure: Repository	Discussion of challenges related to repository platform limitations
Challenges: Infrastructure: Resources	Discussion of challenges related to access to computational resources
Challenges: Infrastructure: Review	Discussion of challenges related to reproducibility review infrastructure
Challenges: Infrastructure: Support	Discussion of challenges related to supporting authors through the process
Challenges: PublicationTime	Discussion of challenges related to publication time
Challenges: Reproducibility	Discussion of challenges in handling cases or irreproducibility
Challenges: ReviewTime	Discussion of challenges related to review time
Challenges: Skills: Author	Discussion of challenges related to author skills
Challenges: UseOfStudents	Discussion of challenges related to use of students
Continuity	Discussions of continuity required by initiative including editorial or conference stakeholders. Includes sustainability
Definitions: Replicability	Interviewee's definition of reproducibility w.r.t. initiative

Code	**Definition**
Definitions: Reproducibility	Interviewee's definition of replicability w.r.t. initiative
Definitions: Transparency	Interviewee's definition of transparency w.r.t initiative
Documents References	Refererence to related documents
Examples	Specific examples, anecdotes
Expertise: Curators	Discussion of expertise required of curators
Expertise: Editors	Discussion of expertise required of editors
Expertise: Reviewers	Discussion of expertise required of reviewers
Expertise: Verifiers	Discussion of expertise required of verifiers
Feedback type	Discussion of type of feedback given (e.g., supportive, corrective, preventative)
Incentives	Discussion of incentives
Key Events	Discussion of key turning events
Leadership	Discussion of role of leadership
Mandate	Discussion of policy mandate
Measurement: Adoption	Discussion of metrics used to measure adoption
Measurement: Behavior	Discussion of ways to measure researcher behavior
Measurement: Citation	Discussion of use of citation-based metrics
Measurement: Cost	Discussion of use of cost-based metrics
Measurement: Downloads	Originally intended to be repository downloads. Expanded to include Github forks/stars.
Measurement: Errors	Discussion of errors as initiative metric
Measurement: Impact	Discussion of impact factors
Measurement: PublicationTime	Discussion of publication time as initiative metric
Measurement: ReviewTime	Discussion of review time as initiative metric
Measurement: Submission	Discussion of submission rates as initaitive metrics
Measurement: Survey	Discussion of use of surveys to measure inititiative
Measurement: UseInClasses	Discussion of use of artifacts in the classroom, i.e., graduate education
Motivation	Underlying motivation of the initiative
Training	Discussion of need or acts of training/education, development of training materials, training reviewers as well as authors
NonReproduction	Discussions related to nonreproductio rejection of submitted artifacts or failure to reproduce or retractions.
Personal encounters	Discussions of personal experiences
Policies	Discussion of guidelines or checklists
Policy Changes	Discussion of changes made to policies, guidelines.
Prestige	Discussion of prestige – journal ranking and type
Protected information	Discussions of private, proprietary, commercial or otherwise protected information.

Code	**Definition**
Qualitative Theory	Discussion of qualitative research or theoretical research
Quality	Discussion of software quality
Reaction: Commmunity	Discussion of author/community reaction
ReproducibilityByWhom	Discussion of who is responsible for actual reproduction or evaluation
ReproducibilityOfWhat	Discussion of what is "reproduced" as
Resources: Human	Discussion of human resources required
Resources: Technical	Discussion of technical or computational resources required
Reuse/usability	Discussion of reuse, extensibility, or generalizability
Risk	Discussion of risks to initiative
Role: Archives	Discussion of role of archives
Role: Commitment	Discussion of the need for or role of organizational buy-in or individual commitment
Role: Education	Discussion of role of reproducibility in education
Scalability	Discussion of scalability of initiative
ScalingUpPilots	Discussion of scaling up initiativeAdded pilots
Workflow: Assignment	Discussion of reviewer assignment in workflow
Workflow: Blindness	Discussion of blindness, peer-review
Workflow: Communication	Discussion of communication process during workflow
Workflow: Documentation	Discussion of documentation including policies and guidelines
Workflow: Infrastructure	Discussion of specific technical infrastructure used in workflow
Workflow: Limitations	Discussion of what the workflow doesn't do (can be broken down)

Table C.2: Detailed codes used for coding of interview transcripts.

C.2 Codes used for documentary evidence

Code	**Definition**
Documentation: README	Guidelines related to the creation of readmes
Documentation: Manifest	Guidelines related to the creation of overall package manifests
Data: Source data	Guidelines related to documenting, citing, attributing, and access instructions for any source datasets used. This includes source data used in the creation of analysis datasets, benchmark datasets, etc.
Data: Data formats	Guidelines related to the format of data (e.g., non-proprietary)
Data: Data license	Guidelines related to data license and copyright including redistribution of data from other sources as well as licenses for author-provided data.
Data: Data documentation	Guidelines related to data documentation including codebooks or other forms of metadata.

Code	**Definition**
Data: Availability	Guidelines related to data availability including embargos, and access protocols unrelated to licensing or proprietary data.
Data: Proprietary	Guidelines related to proprietary, private, confidential, sensitive data.
Data: Sensitive data audit	Guidelines related to audits for sensitive data
Environment: Compiler	Guidelines related to compilers, versions, and related settings including availability.
Environment: Hardware	Guidelines related to required hardware dependencies, state, settings, including availability.
Environment: Software	Guidelines related to software dependencies including applications, libraries, settings, compilation, and availability.
Environment: Runtime	Guidelines related to runtime state or capture of runtime environment information (hot/cold cache)
Environment: Execution conditions	Guidelines related to specification of required execution conditions (single user, process pinning)
Environment: Resource requirements	Guidelines related to specification of resource requirements including disk space, memory, cores, and time required for workflow steps
Environment: External systems	Guidelines related to documentation, availability, and access to external systems.
Environment: Other	
Experiment: Workflow	Guidelines related to documentation of complete experimental workflow including inputs and outputs. This includes specification of workflow framework, where applicable.
Experiment: Results	Guidelines related to documentation of results including figures and tables.
Experiment: Benchmark programs	Guidelines related to documentation of benchmark programs used
Experiment: Metrics	Guidelines related to documentation of metrics reported and used for optimization
Experiment: Evaluation	Guidelines related to experiment evaluation and expected results
Experiment: Customization	Guidelines related to customization of experimental conditions.
Experiment: Parameters	Guidelines related to documenting input parameters
Experiment: Random seed values	Guidelines related to documentation of random seed values
Code: License	Guidelines related to code licenses
Code: Availability	Guidelines related to code availability including archived locations, use of Github or research repositories.
Code: Installation	Guidelines related to installation instructions and installation
Code: Documentation	Guidelines related to code documentation, including comments
Code: Manifest	Guidelines related to documenting relationship between code, data, results (provenance)
Code: Versions	Guidelines related to code versions
Code: Citations	Guidelines related to software citations

Code	**Definition**
Other: Archival formats	Guidelines related to the use of archival formats
Other: Algorithm description	Guidelines related to the specification of algorithms
Other: Integrity checks	Guidelines related to file/data integrity checking
Other: Extensibility	Guildelines related to evaluation of extensibility of systems
Other: Software packaging	Guidelines related to packaging formats
Other: Sensitive data checks	Guidelines related to checking for presense of sensitive data
Other: Link checking	Guidelines related to checking links
Other: Robustness	Guidelines related to checking robustness, i.e., robustness to change.
Other: Observation	Guidelines related to access via observation
Verification: Reproducibility	Guidelines related to the assessment of reproducibility results
Verification: Reproduction	Guidelines related to reproduction and verification of results
Verification: Suitability	Guidelines related to assessment of suitability
Verification: Non-reproduction	Guidelines related to the effect of non-reproduction on publication.
Publishing: Repository	Guidelines related to artifact publication
Publishing: Packaging	Guidelines related to artifact packaging
Publishing: Availability	Guidelines related to artifact availability
Publishing: Naming	Guidelines related to paths and naming
Publishing: Identifers	Guidelines related to identifiers and linking artifacts
Publishing: Authorship	Guidelines related to artifact authorship
Functionality: Statement of need	Guidelines related to describing the need addressed by the artifact
Functionality: Example usage	Gudielines related to providing example usage
Functionality: Functional claims	Guidelines related to functional claims (as opposed to experimental claims)
Functionality: Performance claims	Guildelines related to performance claims
Functionality: Functionality documentation	Guidelines related to documentation of functionality
Functionality: Automated tests	Guidelines related to manual or automated tests
Functionality: Contribution guidelines	Guidelines related to open source contribution documentation

Table C.3: Detailed codes for policies and guidelines.

Appendix D

Documentary Evidence

Table D.1 includes a complete listing of the documentary evidence used in the qualitative analysis.

Initiative	**Documents**
AJPS	AJPS Verification Policy Replication and Verification Policy Guidelines for Preparing Replication Files AJPS Dataverse Quantitative Data Verification Checklist Qualitative Data Verification Checklist 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) Anti-DART Petition AJPS Editorial Reports 2012-2019 Should Journals Be Responsible for Reproducibility? [135] Verification Verification Our Experience with the AJPS Transparency and Verification Process for Qualitative Research Celebrating Verification, Replication, and Qualitative Research Methods at the AJPS Some Details about New AJPS Submission Requirements QDR and the AJPS Replication Policy AJPS to Award COS Open Practice Badges
AEA	Data and Code Availability Policy AEA Data and Code Repository

Initiative	Documents
	Guidance on how to deposit data at the AEA Data and Code Repository Data and Code Availability Policy: Frequently Asked Questions Verification guidance Example replication report Training and Guidance for assessing replicability Unofficial guidance on various topics by the AEA Data Editor Report by the AEA Data Editor [267] Updated AEA Data and Code Availability Policy (July 16, 2019) Reproducibility and Replicability in Economics Workflow Job posting
JASA	Reviewer Guidelines (via Email) JASA-ACS Github organization Reproducible Research in JASA JASA Editors Talk Reproducibility Author Contributions Checklist form Author Instructions
IS	Invited Reproducibility Papers - Author Guidelines Invited Reproducibility Papers - Reviewer Guidelines Guide for Authors A collaborative approach to computational reproducibility New article type verifies experimental reproducibility
Biostatistics	Information for Authors Reproducible research and Biostatistics [211] Editorial [76] Reproducible research and the substantive context [141] Discussion of Keiding [212] Reproducible research and the substantive context: response to comments [142]
TOMS	The TOMS Initiative and Policies for Replicated Computational Results (RCR) Editorial: ACM TOMS Replicated Computational Results Initiative [122] RCR Reviewer Invitation (via Email)

Initiative	**Documents**
SC	SC 19 Reproducibility Initiative From SC Papers to Student Cluster Competition Benchmarks: Joining Forces to Promote Reproducibility in HPC AD/AE Appendices Track Report SC19 process AD/AE Appendices Author FAQ Paper submissions Appendix Review Instructions Reproducibility Challenge Track Journal Special Issue Track SC Reproducibility Materials Student Cluster Competition Student cluster competition: a multi-disciplinary undergraduate HPC educational tool Parallel Computing special issue (SC16) Special Issue on SCC?17 Reproducibility Initiative Special Issue on the SC?18 Student Cluster Competition Reproducibility Initiative

Table D.1: Documentary evidence used in qualitative coding

Appendix E

Workflows

E.1 American Journal of Political Science

This section describes the verification workflow for quantitative research as implemented by AJPS and Odum.

E.1.1 Data sources

1. Interview transcripts
2. AJPS Email Templates Examples.docx
3. Curation + Verification Workflow.pdf
4. Journals_CurationChecklist.docx
5. Journals_CurationProcedures_Current.txt
6. Journals_VerificationChecklist.docx
7. JournalVerifier_NDA.docx
8. Verification_ResultDefinitions.pdf
9. VM_Instructions_Verifier.docx

E.1.2 Workflow summary

Prerequisites

1. Odum verifiers are required to sign a Non Disclosure Agreement (NDA) for non-disclosure of confidential information.
2. Odum curators have administrative privileges on the AJPS Dataverse
3. Author submits manuscript via Editorial Manager

Editorial review

- AJPS editorial team reviews manuscript using established editorial process. This includes desk rejection, peer review (double-blind), and revise & resubmit (11R&R"). This process is managed using the Editorial Manager system.
- Manuscript is conditionally accepted and authors are notified of the requirement to submit replication materials to the AJPS Dataverse. Guidelines are provided on the ajps.org website under “Guidelines for Accepted Articles" including references to:
  - Guidelines for Preparing Replication Files
  - Quick Reference for Uploading Replication Files

  - Quantitative Data Verification Checklist
  - Qualitative Data Verification Checklist

- Authors are given 2 weeks to upload replication materials to the AJPS Dataverse and notify the editor via the “Submit for Review" button in Dataverse.
- The managing editor notifies the Odum Archive that the materials are available (via shared email account in Outlook) attaching the manuscript draft. The managing editor enters a record for the manuscript in the tracking spreadsheet.
- Note: The Odum verification process determines article publication, but is otherwise not considered part of the peer review process.

Curation

- For new submissions, curators assign the manuscript to the next curator in rotation using Outlook labels. For resubmissions, the previously assigned curator remains assigned except under unusual circumstances. Curators act as the central point of contact between the Archive and AJPS (i.e., verifiers never communicate with editors or authors).
- Curator follows the curation workflow (VeriWorfklowDetailed.pdf)
  - Creates manuscript folder on Odum Archive shared filesystem
  - Copies manuscript to folder
  - Downloads replication materials from Dataverse (downloads as zip) to Datafiles folder.
  - Creates new record in Dashboard database (Manuscript number, author, Dataverse DOI)
  - Runs md5checker to capture checksums at point of submission, enters into Dashboard.
  - Reviews materials based on Data Curation Checklist
  - For restricted access data, tries to obtain a copy
  - Updates Dashboard curation record (date, time taken, curator)
  - Writes curation report
  - If materials are complete: assigns verifier (see below). If materials are incomplete: sends email to managing editor via archive account attaching curation report. In Dataverse, returns package to author.

Verification

- If new submission, verifiers are assigned based on rotation, load, and expertise under some circumstances. If resubmission, prior verifier remains assigned except under unusual circumstances. Assignment is made via the Odum Dashboard database and email.
- Verifier follows verification workflow:
  - Sets up environment (based on VM_Instructions_Verifier.docx)
  - Enters Dashboard information (verification date, verifier name, software version, time spent, verification result, verification report)

- Verifier processes vary but include:
  - Downloading a copy of the replication materials and manuscript to VM.
  - Skim manuscript – typically abstract, methods, findings, and discussion – looking for empirical results. AJPS compiles tables and figures at the end. Highlight where there are in-text results.
  - Optionally inspect README and datasets (typically rely on curation workflow)
  - Read the README file and follow instructions
    - Install packages
    - Run scripts, automating if possible. Identify result outputs (console or file) and compare to manuscript
    - Tables: Verify output against manuscript
    - Figures: Visual inspection against manuscript
    - In-text: Identify empirical results in manuscript or via comments.
    - Handle secure data where required
    - Troubleshoot errors/discrepancies
    - Write verification report
  - Under difficult circumstances, verifiers reach out to Odum research services, cyberinfrastructure, and UNC research computing for additional support.
  - Update Dashboard with verification result and report
  - Upon completion, notify curator
- On success:
  - Manuscript moved to “Completed" folder on Archive filesystem.
  - If private/proprietary data, files are deleted after verification.
  - Curator updates Dataverse metadata: assign badges, adds verification note, adds terms of use (private/proprietary), adds data citations
  - Publishes dataset
  - Notifies editor
- On failure: email AJPS editor with verification report; return dataset in dataverse; begin resubmission process
- Curator emails AJPS (AJPS Email Templates Examples.docx): Result - Success; Result - Success with Modifications; Result - Issues (major)
- Managing editor notifies publisher that manuscript is accepted. Link manuscript to Dataverse DOI.

E.2 Supercomputing

This workflow description is based on SC19. The SC reproducibility initiative is continually evolving and additional changes are anticipated. The SC Reproducibility Initiative has four distinct components:

1. Artifact Description / Artifact Evaluation (AE/AD) committee reviews appendices submitted with all technical program submissions (mandatory in 2019)
2. SCC Reproducibility Challenge committee reviews appendices for suitability for SCC and selects 3 for in-person interviews at SC. One paper is selected for SCC RC.
3. SCC RC preparation and actual challenge at SC+1. Students write reports and, in the past, top performers have been selected to be published in a special issue of *Parallel Computing*. The original paper receives award and gets ACM Replicated badge.
4. Publication of Parallel Computing special issue.

E.2.1 Artifact Description/Artifact Evaluation

- Paper authors submit manuscripts via Linklings submission interface that, as of 2019, includes AD/AE fields (see Appendix E).
- AD/AE Appendix Committee reviews appendices submitted with all technical papers. This is a double-open process managed partially via Linklings and partially using custom software / email.
- Reviewers follow the “Appendix Review Instructions" and assign Artifact Available badge where merited
  - AD/AE appendices are reviewed for completeness. A complete AD appendix is required. Artifacts are not required to be available. Therefore, the reproducibility of a submitted article is assessed with no badge assigned.
  - Criteria for incompleteness include used when they were
  - Summary of experiments references text of the paper or does not enumerate artifacts and their purpose
  - URL/DOI is broken or leads to a resource without substance
  - Version information is missing from hardware, OS, compiler etc.
  - If fields are left blank but relevant artifacts are indicated in the text of the paper
  - If modifications are not specified but are included in the text of the paper
- Papers with AD appendices are processed by Reproducibility Challenge (RC) committee[1])

E.2.2 SCC/Reproducibility Challenge

- Reproducibility Challenge committee chairs request access to Linklings for all papers with AD appendix from SC-1 (there were 40 in 2018)
- RC committee reviews papers and sets up 3 interviews during SC0.

[1] See `https://github.com/SC-Tech-Program/SCreproducibility/blob/master/Reproducibility-Challenge.md`

  - 1) Feasibility for competition, one reviewer per paper e.g., use of Summit, architecture specific, suitability
  - 2) Ideal application for SCC
    * Deep read
    * How do we convert this to something an undergraduate could do in 2-3 months (when it was grad students for much longer)
  - From interview – 3 criteria:
    * Openness – open-source and available, no proprietary dependencies; specific to some hardware; feasibility
    * Applicability – to the scale for students
    * Science story – has to get undergraduates excited about HPC
    * Tie-in to technical program (e.g., ML/DL for next year)
  - Committee rates them and gives overall score; meet as a committee and select top 3 the interview at SC0.

- Author interviews at SC0
  - See interview questions in section below
  - Requires commitment, potential modification, support of students. Need expertise of authors (e.g, to get code that ran on Titan or Stampede2 to run on student cluster)

- SC+1 RC committee selects paper (January/March) and announces in blog post.
- SC+1 RC committee creates the challenge
  - Committee members run the application, make sure things go as expected – center staff and faculty. Get authors to fix/change where possible.
  - Organizers must know the outcome to grade students

- Send exercise to SC+1 participants (March)
  - Coordinate webinars

- Develop grading rubric
- SC+1 (November)
  - SC+1 – give new dataset that wasn't in the paper
  - Students participate in SCC competition

- Students write reports, reviewed by SCC, score them. Declare winner
- Original article from SC-1 receives badge - Artifact Available, Artifacts Evaluated - Functional, and Results Replicated. Author receives certificate of appreciation at SC+1
- SC+2
  - Student reports are published prior to the conference so that they can be shared.

E.2.3 Parallel Computing special issue

Top X papers are selected and shepherded through publishing process for special issue of Parallel Computing e.g., https://www.sciencedirect.com/journal/parallel-computing/vol/90/suppl/C

E.2.4 RC Author Interview

Questions for the author team[2]:

1. What platforms and architectures can this application run on?
2. Do you expect similar or identical results on a different hardware architecture?
3. How much data does the application require for input? How much data does it produce?
4. Can you produce multiple (at least 2) scaled down data sets to run on student clusters?
5. Can the application be scaled to run on a modest-sized cluster?
6. Can your team commit the time to working with the SCC team and the Reproducibility team to refactor your application/dataset and work with the student teams to get the code working?

Questions for the Reproducibility team to consider:

1. We need to gauge the author team's excitement, interest, and whether they?ll be able to devote the time
2. Does the application lend itself to a science story that can excite both the student teams and the public? Bonus points if it aligns with an SC theme or local/regional tie-in

E.3 American Economic Association

This section describes the verification workflow for AEA papers as implemented by the Data Editor at LDI.

E.3.1 Data sources

1. Interview transcripts
2. Workflow
3. Verification guidance
4. Example replication report
5. Training and Guidance for assessing replicability

[2] From `https://github.com/SC-Tech-Program/SCreproducibility/blob/master/Reproducibility-Challenge.md`

E.3.2 Workflow summary

Prerequisites

1. Data Editor and "approvers" have administrative permissions on AEA OpenICPSR
2. Verifiers must have access to OpenICSPR and Atlassian tools (JIRA, Bitbucket) and Cornell netid to access compute resources.

Manuscript submission

1. Author prepares and submits materials to a supported repository based on AEA policy.
2. Author submits manuscript via ScholarOne and acknowledges Data and Code Availability policy.

Editorial review

1. AEA journal editorial team reviews the manuscript based on established editorial processes. This includes desk rejection, peer review, and revise & resubmit. This process is managed using ScholarOne.
2. Upon conditional accept, manuscript files are compiled and the AEA Data Editor is assigned for review via ScholarOne. An email is sent and via JIRA integration a ticket is created.

Data review and verification

1. JIRA issue is assigned to a verifier who follows the verification guidelines
2. Creates repo in private Bitbucket and populates with manuscript, readme, replication report template, replication package.
3. Fills out "Data citation and information report"
4. Checks for personally identifiable information
5. Writes replication report
6. Approver reviews report and submits via ScholarOne (as data editor)

E.4 Biostatistics

This section describes the workflow for the verification of reproducibility of materials submitted to the journal *Biostatistics*.

- Authors submit manuscript and optional supplementary materials via ScholarOne.
- Upon acceptance, authors were notified in a letter from the editor that they could volunteer for the reproducibility review.
- If data and code are provided and code was written in R, it was eligible for the AER review and resulting kitemark.
- The manuscript was assigned to the AER for review. The AER ran the provided code and confirmed that outputs matched those reported in the paper.

- Communication with the author was within ScholarOne
- Upon successful review, the paper received the R kitemark and was published.

Notes:

1. For data and code kitemarks, there was no review at all (honor system)
2. There was no procedure in the event of non-reproduction

Sources:

1. Interview transcripts
2. “Reproducible research and Biostatistics" [211]
3. *Biostatistics* Information for Authors

E.5 Information Systems

This section describes the workflow for initiative as implemented by *IS*.

- IS journal editors identify a candidate reproducibility paper (criteria unknown). Authors are invited by the editors to submit a reproducibility paper. Reproducibility papers must follow the same guidelines as regular papers.
- Authors follow the published guidelines to submit the invited paper via Editorial Manager and materials via Mendeley Data.
  - Authors must provide source code for software components and installation scripts or a URL and version for a repository or hosting service. Authors are encouraged to submit virtual machine images, ReproZip RPZs or Docker images.
  - Authors submit a reproducibility article with the following information
    - Details about the computational environment
    - Explanations about different data and input parameters
    - Instructions for installing and compiling software
    - Instructions for running experiments and producing plots and tables
    - Limitations, if any.
  - Authors publish software and data to Mendeley Data as a single dataset
  - Reviewers are recruited to reproduce the results in the paper (double-open). They verify results presented in the paper can be reproduced and test how robust the results are to changes in the experiment configuration
  - Upon acceptance, the paper is published in the “Reproducibility Section" of the journal with original authors and reviewers as co-authors.

E.6 Journal of the American Statistical Association

This section summarizes the operational workflow based on interview transcripts:

- An author submits a manuscript using ScholarOne including required ACC Form and optionally code/data as supplemental information.
- ACC Form and other supplemental materials are deposited in Figshare
- The manuscript goes through regular peer review process
- If paper is “on a path to acceptance" (i.e. conditional accept), it is assigned to the coordinating AER via ScholarOne using a fake reviewer account (“Dr. Reviewer for Reproducibility")
- The coordinating AER assigns an reviewing AER via round-robin
- The assigned AER reviews ACC Form and materials according to Guidelines and Evaluation Criteria. AERs are required to "assess the completeness and quality of the ACC form and provide a general assessment of whether the results could be reproduced based on the artifacts provided" and may optionally run the code.
- AER reviews are returned via ScholarOne. Upon success, manuscripts are assigned the “Reproducibility materials accepted" decision and accepted for publication.
- Once the manuscript is accepted, it is published online with supplementary materials in Figshare. The AER begins the process of copying materials from external repositories into the JASA-ACS Github.

Sources: Interview

E.7 Transactions on Mathematical Software

The workflow for the *TOMS* RCR process is detailed on the journal website and has been summarized here.

- Authors submit manuscript, requesting optional RCR review
- Paper goes through standard review process
- Paper is reviewed for suitability for RCR
- RCR reviewer is recruited and assigned
- RCR reviewer is responsible for replicating the reported results
- RCR reviewer documents details of replication process
- RCR reviewer provides determination.
  - On failure, manuscript is not accepted and may be returned to author.
  - On success, paper and RCR report are published. RCR referee is acknowledged on the primary paper.

The journal relies on the expertise of the reviewer to determine the process for replication and make the final determination. The policy describes two modes of replication:

1. Independent replication: The authors provide the RCR reviewer access to, or sufficient description of, the computational platform used to produce the manuscript results. Access could be: A direct transfer of software to the reviewer or a pointer to an archive of the software, and a description of a commonly available computer system the reviewer can access. A guest account and access to the software on the system used to produce the results. Detailed observation of the authors replicating the results.
2. Review of computational results artifacts: In some situations, authors may not be able to readily replicate computational results. Results may be from a system that is no longer available, or may be on a leadership class computing system to which access is very limited. In these situations, careful documentation of the process used to produce results could be sufficient for an RCR designation. In this case, the software should have its own substantial verification process to give the reviewer confidence that computations were performed correctly. If timing results are reported, the authors' artifacts should include validation testing of the timers used to report results.

Appendix F

Artifacts

This appendix lists the artifacts reviewed in the analysis presented in Chapter 7. The five most recently published papers from each initiative were selected. For initiatives with fewer than five verified papers, all papers were selected.

AEA

1. Bernanke, Ben. Data and Code for: “The New Tools of Monetary Policy." Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-03-10. https://doi.org/10.3886/E117206V1
2. Bach, Laurent, Calvet, Laurent, and Sodini, Paolo. “Rich Pickings? Risk, Return, and Skill in Household Wealth." Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-03-10. https://doi.org/10.3886/E117466V3
3. Farboodi, Maryam, and Veldkamp, Laura. “Data and Code For: Long Run Growth of Financial Data Technology." Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-03-05. https://doi.org/10.3886/E114984V2
4. Elder, Todd, and Zhou, Yuqing. “Analysis Code for The Black-White Gap in Non-Cognitive Skills among Elementary School Children." Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-03-04. https://doi.org/10.3886/E117301V1
5. Bhandari, Anmol, Birinci, Serdar, McGrattan, Ellen R., and See, Kurt. “Data and Code for: What Do Survey Data Tell Us about US Businesses" Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-03-03. https://doi.org/10.3886/E117021V3

AJPS

1. Casas, Andreu, Matthew J. Denny, and John Wilkerson. 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. Paper: doi:10.1111/ajps.12472, Data: doi:10.7910/DVN/7ZVSYO.
2. Brierley, Sarah, Eric Kramon, and George Kwaku Ofosu. 2020. “The Moderating Effect of Debates on Political Attitudes." American Journal of Political Science 64(1): 19-37. Paper: doi: 10.1111/ajps.12458, Data: doi:10.7910/DVN/OJA7YS.
3. Haynes, Kyle, and Brandon K. Yoder. 2020. “Offsetting Uncertainty: Reassurance with Two-Sided Incomplete Information." American Journal of Political Science 64(1): 38-51. Paper: doi: 10.1111/ajps.12464, Data: doi: 10.7910/DVN/PXOT5L.

4. Nielsen, Richard A. 2020. “Women?s Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers." American Journal of Political Science 64(1): 52-66. Paper: doi: 10.1111/ajps.12459, data:, doi:10.7910/DVN/6YNZTE.
5. Strickland, James M. 2020. “The Declining Value of Revolving-Door Lobbyists: Evidence from the American States." American Journal of Political Science 64(1): 67-81. Paper: doi:10.1111/ajps.12485, Data: doi:10.7910/DVN/YQYZ6O

Biostatistics Materials for these articles are no longer available.

1. Lee, Duncan, Claire Ferguson, and Richard Mitchell. 2009. “Air Pollution and Health in Scotland: A Multicity Study." Biostatistics 10(3): 409?23.
2. Magi, Alberto et al. 2010. “A Shifting Level Model Algorithm That Identifies Aberrations in Array-CGH Data." Biostatistics 11(2): 265?80.
3. Magi, Alberto et al. 2010. “A Shifting Level Model Algorithm That Identifies Aberrations in Array-CGH Data." Biostatistics 11(2): 265?80.
4. Riebler, Andrea, and Leonhard Held. 2010. “The Analysis of Heterogeneous Time Trends in Multivariate Age?Period?Cohort Models." Biostatistics 11(1): 57?69.
5. Varin, Cristiano, and Claudia Czado. 2010. “A Mixed Autoregressive Probit Model for Ordinal Longitudinal Data." Biostatistics 11(1): 127?38.

Information Systems

1. Wolke, Andreas, Martin Bichler, Fernando Chirigati, and Victoria Steeves. 2016. “Reproducible Experiments on Dynamic Resource Allocation in Cloud Data Centers." Information Systems 59: 98?101. doi: 10.1016/j.is.2015.12.004
2. Lastra-Diaz, Juan J. et al. 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. doi:10.1016/j.is.2017.02.002
3. Farina, Antonio et al. 2019. “On the Reproducibility of Experiments of Indexing Repetitive Document Collections." Information Systems 83: 181?94. doi:10.1016/j.is.2019.03.007

JASA-ACS

1. Banerjee, Trambak, Gourab Mukherjee, Shantanu Dutta, and Pulak Ghosh. 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 0(0): 1-29. doi: 10.1080/01621459.2019.1611584
2. Lee, Clement, and Darren J. Wilkinson. 2020. “A Hierarchical Model of Nonhomogeneous Poisson Processes for Twitter Retweets." Journal of the American Statistical Association 115(529): 1-15. doi: 10.1080/01621459.2019.1585358
3. Smith, Adam N., and Greg M. Allenby. 2020. “Demand Models With Random Partitions." Journal of the American Statistical Association 115(529): 47-65. doi: 10.1080/01621459.2019.1604360

4. Tang, Xueying et al. 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): 1561-73. doi: 10.1080/01621459.2019.1585250
5. Wilson, Douglas R., Chong Jin, Joseph G. Ibrahim, and Wei Sun. 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 0(0): 1-11. doi: 10.1080/01621459.2019.1654874

Supercomputing

1. Ben-Nun, Tal et al. 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, SC19, New York, NY, USA: Association for Computing Machinery. doi:10.1145/3295500.3356173.
2. Domke, Jens 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, New York, NY, USA: Association for Computing Machinery. doi:10.1145/3295500.3356140.
3. Laguna, Ignacio 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, New York, NY, USA: Association for Computing Machinery. doi:10.1145/3295500.3356176.
4. Li, Lingda, and Barbara Chapman. 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, New York, NY, USA: Association for Computing Machinery. doi:10.1145/3295500.3356141.
5. Narra, Krishna Giri 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, New York, NY, USA: Association for Computing Machinery. doi:10.1145/3295500.3356170.

TOMS

1. Replicated Computational Results (RCR) Report for BLIS: A Framework for Rapidly Instantiating BLAS Functionality
2. Replicated Computational Results (RCR) Report for A Sparse Symmetric Indefinite Direct Solver for GPU Architectures
3. Replicated Computational Results (RCR) Report for A Distributed-Memory Package for Dense Hierarchically Semi-Separable Matrix Computations Using Randomization
4. Replicated Computational Results (RCR) Report for Code Generation for Generally Mapped Finite Elements