Series Name: WPS
Paper No.: 249
Issue Date: July 31, 2025

Assessing Reproducibility in Economics Using Standardized Crowd-sourced Analysis

Abel Brodeur, Seung Yong Sung, Edward Miguel, Lars Vilhuber, Fernando Hoces de la Guardia

CEGA

Center for Effective Global Action

Working Paper Series

Center for Effective Global Action
University of California

Assessing Reproducibility in Economics Using Standardized Crowd-sourced Analysis*

Abel Brodeur, University of Ottawa and Institute for Replication
Seung Yong Sung, University of California, Berkeley
Edward Miguel, University of California, Berkeley and NBER
Lars Vilhuber, Cornell University
Fernando Hoces de la Guardia, University of California, Berkeley
ⓡ†

April 2025

Abstract

This paper presents a framework to standardize crowd-sourced computational reproductions in economics through the Social Science Reproduction Platform (SSRP). The approach address four main challenges for computational reproductions: a lack of standardization, aggregation issues, existing incentives for "adversarial" interactions, and the loss of knowledge from analyses that are never published. We then summarize the first 487 reproductions uploaded on the SSRP. The results show substantial heterogeneity in the ability to successfully reproduce empirical results in economics research, with approximately 30% of recent studies meeting at least a basic definition of being computationally reproducible.

Keywords: Reproducibility, Replication, Open Science, Research Transparency, Social Science Reproduction Platform
JEL codes: A14, C10, C81

*Corresponding author: fhoces@berkeley.edu. We thank Joel Ferguson, Alex Bogdanoski, and Albert Chae, Wenjing Lin for providing valuable feedback and assistance on this project. Funding for the SSRP was provided by Arnold Ventures and the Berkeley Initiative for Transparency in the Social Sciences. All errors are our own.

†ⓡ Denotes random ordering of authors. This ordering was generated by the AEA randomization tool and is archived under confirmation code m3VgFI27yjgY.

1 Introduction

The question of research reproducibility is at the forefront of current scientific debates. Computational reproducibility – the ability to produce "consistent results using the same input data, computational methods, and analytical conditions" (National Academies of Sciences, Engineering, and Medicine 2019) – is tightly linked to some of the most important principles of science.[1] Two of these principles, as articulated in Merton (1942)'s seminal discussion, are that of *communality* and *organized skepticism*. Communality defines science as a collaborative problem-solving enterprise, in which intellectual products are freely shared in ways that facilitate collaboration, discussion, and reuse. Organized skepticism specifies that all scientific claims are critically examined and evaluated by the scientific community before being accepted as true. Both communality and organized skepticism are actualized through computational reproducibility (or simply "reproducibility" in what follows).

Recent work assessing the current state of reproducibility in economics using varying approaches and definitions of reproducibility (Chang and Li 2022, Gertler et al. 2018, Wood et al. 2018) finds that only between 14% and 43% of published studies were computationally reproducible.[2] Reproducibility appears to fail for many reasons, including unavailable or incomplete computer code (Gertler et al. 2018), incomplete documentation, versioning issues for software and packages, and the numerical fragility of some empirical results (Chang and Li 2022). The limited numbers of published reproductions and low rates of reproducibility are perhaps unsurprising in light of the many barriers that prevent researchers from assessing the reproducibility of existing research and the limited number of journals enforcing their data and code availability policy in practice (Brodeur et al. 2024a). Reproducibility depends on researchers communicating clear and complete accounts of their data, computer code, methods, results, and associated uncertainties so that the scientific community can accurately evaluate the findings in question. This process allows researchers to "build on others' work, develop the necessary skills to conduct high-quality studies, and check results and confirm, dispute, or refine them" (National Academies of Sciences, Engineering, and Medicine 2019).

While the growing literature noted above and other recent work documents the prevalence and extent of reproducibility and replicability in economics and other disciplines (Ankel-Peters et al. 2023, Berry et al. 2017, Camerer et al. 2016, 2018, Hamermesh 2017, Huntington-Klein et al. 2021, Sukhtankar 2017, Vilhuber 2020, Brodeur et al. 2024b), the question of how to standardize and aggregate reproduction exercises has not yet received sustained attention. This paper aims to begin to address this gap. In particular, we describe a new tool to address leading concerns regarding computational reproducibility in economics, the Social Science Reproduction Platform (SSRP),

[1] In this paper we aim to follow closely the following NAS definitions of reproducibility and replicability:

> "[...] reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis. This definition is synonymous with "computational reproducibility," and the terms are used interchangeably [...].
>
> Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." (National Academies of Sciences, Engineering, and Medicine 2019, p.46)

In economics, replications are also understood to be some variation of the original analysis using the same data (a new robustness check or sensitivity analysis), blurring the line of where a reproduction ends and a replication begins. Most of this paper focuses on reproducibility but in the last section we discuss how to accommodate robustness checks and improvements in the original analysis.

[2] In response to these issues, data editors at several economics journals have begun coordinating pre-publication checks for computational reproducibility to ensure that the data and code published alongside the research articles successfully reproduce the results presented in the paper (see American Economic Association (AEA) 2020 for a pioneering effort). However, the practice remains far from universal, especially outside of the most cited journals.

which was developed by the authors of this article (and others) and provides a standardized approach to performing and recording reproductions.[3]

We first provide a description of the SSRP and how it may help alleviate many of the leading challenges faced by those carrying out reproductions in economics and other social sciences, and lay out potential use cases. We then present descriptive statistics and analysis based on the first 487 reproductions submitted to date, with a particular emphasis on the 411 reproductions that have been completed in economics (while the remainder to date have mainly been carried out in political science).[4] We find that papers reproduced in the SSRP are more recent and more cited than a broadly representative sample of articles published in the top 100 Economics journals since 2010. A main finding of the analysis is that the reproducibility levels of papers in economics, reproduced on SSRP, show substantial variation; 14% of such papers were successfully reproduced from the raw project data, and another 16% reproduced only from the analysis data (i.e., data that has already been processed, for instance, to create key variables). Taken together, and in the main empirical finding of this paper, these figures indicate that in roughly 30% of the papers examined so far all key results examined meet at least some definition of being reproducible.

Whether this is a surprisingly high or low number depends on one's priors, but regardless of whether one views these optimistically or pessimistically, it is clear that a large share — roughly 70% — of recently published economics paper have at least some results cannot be readily computationally reproduced. These figures differ somewhat depending on whether one examines research claims separately or aggregates up to an assessment at the paper level, but in either case the broad patterns remain the same. This may in fact be overly optimistic if the papers that are the focus of crowd-sourced efforts on SSRP are more likely to be reproducible.

The reproducibility rates reported paint a picture broadly consistent with the relatively low assessments mentioned above in Chang and Li (2022), Gertler et al. (2018) and Wood et al. (2018), but contrast with Brodeur et al. (2024b)'s assessment of 110 recently published articles. They find a high rate of fully computationally reproducible results using the data and code provided by the authors (over 85%). Our work differs from Brodeur et al. (2024b) as they mostly target journals internally conducting computational reproducibility checks and their efforts emphasize sensitivity analysis.

These results contribute to the small but growing literature discussing the credibility of findings in economics, and more generally the social sciences (Andrews and Kasy 2019, Brodeur et al. 2020, Ioannidis et al. 2017, McCullough et al. 2008, Swanson et al. 2020). We also see this work as contributing to and complementing existing tools for making economics and social science research more credible and replicable, including editorial statements (Blanco-Perez and Brodeur 2020), pre-registration and pre-analysis plans (Banerjee et al. 2020, Coffman and Niederle 2015, Olken 2015), replication markets, and other journal-based policies (Brandon and List 2015, Camerer et al. 2019, Drazen et al. 2021).[5] Yet without computational reproducibility, it is not possible to critically

[3]The development of the SSRP was funded by Arnold Ventures and the Berkeley Initiative for Transparency in the Social Sciences (BITSS), with intellectual support from the AEA Data Editor (Vilhuber).

[4]Data in this version of the paper includes SSRP reproductions through March 06, 2025.

[5]The scientific community has provided multiple constructive responses to the problem of low reproducibility. Several centers and initiatives have emerged in the last 15 years to articulate and promote tools and practices that aim at increasing reproducibility (e.g., Berkeley Initiative for Transparency in the Social Sciences (BITSS, bitss.org), the Center for Open Science (COS, cos.io), and Institute for Replication (I4R, i4replication.org)). New services have also emerged to facilitate and/or verify computational reproducibility by third parties. For example, CodeOcean (codeocean.com) and WholeTale (wholetale.org) provide computing capsule services to share an entire computing environment, while Cascad (cascad.tech) provides the services of certification by an external reviewer. This is a dynamic area and new efforts have been emerging every year.

evaluate scientific publications and build on previous work, so we view it as foundational to the rest of the open science enterprise. This point, originally laid out in Buckheit and Donoho (1995), Donoho et al. (2009), implies that measuring and providing tools to increase the computational reproducibility of research findings is of critical importance to any research discipline.

A further potential contribution is that the SSRP will allow the work that thousands of graduate students and researchers around the world perform each year when they carry out computational reproductions as part of their coursework (or independent research) to be recognized and made publicly available to the research community. In the absence of a centralized platform like SSRP, the knowledge generated by these exercises is largely lost to the community. This is a big loss for science and potentially society as a whole, since discovering instances where findings do not reproduce is valuable for scientific progress and public policy choices. Verifying reproducibility also has significant cost implications. For example, Colliard et al. (2022) estimate that reproductions in ten leading economics journals cost an average of USD $365 per research article. The budget for the AEA's data editor activities accounts for around 5% of the operation costs of its 9 journals (Cherry Bekaert LLP 2022), which, if allocated per paper (415 papers in 2021, Vilhuber 2022) amounts to about USD $750 per paper.[6] The crowdsourcing of computational reproductions from the community can help to fill in these gaps at low cost to professional organizations and journals, and may provide greater coverage than is currently offered by data editor services.

The rest of the paper is structured as follows. Section 2 provides an overview of patterns of reproducibility in economics. Section 3 describes the SSRP, and Section 4 then investigates the characteristics of papers reproduced on the SSRP and provides reproducibility rates for the first 411 reproductions completed for economics articles. The final section concludes.

2 Challenges to Reproducibility in Economics

There are, at least, four key challenges to computational reproducibility that the SSRP aims to address (at least partially), in addition to the low reproducibility rates discussed above.

First, computational reproducibility exercises can be carried out in many different non-standardized ways. One reproduction attempt could dedicate a few hours to verify only the main results of the paper while another could dedicate several days to thoroughly check the entire paper and extend certain secondary analyses. As a result, it is challenging to compare different reproductions of a given paper, and the lack of standardization could readily lead to different and conflicting assessments.

Second, the lack of standardization hinders the aggregation of reproduction exercises that are already conducted every year around the world. Each semester, a large number of graduate courses assign a computational reproduction exercise of a published paper, and the students completing these assignments generate valuable scientific knowledge. This knowledge can be broadly grouped into four key steps: (i) the student reproducers verify if the paper has any publicly available reproduction materials; (ii) they assess the degree of reproducibility of an entire paper or a subset of its main claims; (iii) they may improve on the current version of the reproduction materials (by finding and fixing errors, for instance), and (iv) they can generate valuable new robustness checks and extensions. The knowledge generated, however, is largely lost at the end of the course exercise as the resulting research product does not fit neatly into existing journal article formats, and there has not been a standardized way of recording the assessments. This represents a substantial degree of wasted

[6]The true cost may be somewhat lower since the AEA data editor budget covers activities that go beyond the narrow computational verification of articles, but this is a reasonable starting point.

research effort for the field, not to mention that the errors discovered are often not corrected in the scientific record.

Third, the reproduction exercises that do attempt to get published have often gravitated toward adversarial exchanges between the reproducers and original authors (Janz and Freese 2019). There is a perception that original authors, who are on average more senior and well-established (Gertler et al. 2018), may sometimes use their status to deter potentially critical reproduction results. At the same time, the reproduction authors – if they wish to successfully get their work published – have strong professional incentives to emphasize those estimates that go "against" the original paper's main results rather than providing a more even-handed assessment. The lack of standardization in computational reproduction exercises exacerbates these concerns about selective and tendentious reporting of results by both the reproducers and original authors.

Fourth, and as already noted, the outputs of reproduction exercises often do not fit neatly into existing journal article formats, so few end up in the public research record. While there have been some recent efforts to overcome this concern, none have become widely adopted in economics.[7]

3 Description of the SSRP

Here we describe how the SSRP provides a framework for assessing and increasing the reproducibility of research findings in economics and other disciplines. We discuss the conceptual motivation behind the SSRP and the structure of its current web presence, and explain how the SSRP's features help to address the four interrelated issues highlighted in the previous section (i.e., the lack of standardization, difficulties in aggregation, adversarial interactions, and the loss of scientific knowledge).

3.1 Conceptual Framework

The SSRP was developed to provide a standardized framework to carry out reproduction exercises, filling in an existing gap in economics. The framework is accompanied by guiding documentation (BITSS, 2022) that breaks down a reproduction into 5 main stages: (i) selection of the paper to be reproduced, (ii) defining the scope of the reproduction exercise, (iii) assessment of reproducibility, (iv) description of any improvements to the analysis or documentation, and (v) description of any new robustness checks. It also provides guidance on how to facilitate constructive communication between authors on a reproduction, and how to build and grade reproduction exercises in the classroom[8] (Figure A.8 in the Appendix presents a screenshot of the guide).

The SSRP's approach attempts to provide a more nuanced assessment of what it means for results to be computational reproducible, with two main departures from existing approaches. First, instead of focusing on the reproducibility of a paper as a whole, the SSRP encourages reproducing authors to focus on specific display items – i.e., tables, figures, inline estimates – and assesses each individually.

Second, rather than taking a dichotomous approach — classifying a display item or a paper as either reproducible or not — the SSRP has developed a 10-point reproducibility scale that captures the substantial variation and "gray areas" in reproducibility attempts. Table 1 contains a summary of the approach and a description of each level of reproducibility. In the SSRP scale, 1 denotes the inability to reproduce the results at all since no data or code are available. Levels 2, 3 and

[7]The best known attempt is probably the Replication Wiki, established in 2014, which links the title of a reproduction written up in a paper to the title of the original paper. Yet over a decade of existence, only around 59 reproductions (defined in the site as "Replication in the narrow sense" have been linked (as of September 2024).

[8]Full guide available at: https://bitss.github.io/ACRE/

4 represent increasing degrees of access to analysis materials without achieving computational reproducibility. A 5 on the scale denotes a result that is reproducible from the analysis data, or **CRA (Computationally Reproducible from Analytic Data)** but not from the raw data (for instance, in cases where the raw data are not publicly available).

Levels 6, 7, and 8 involve the availability of additional materials (e.g., some raw data or code for cleaning raw data) beyond the analysis code and data, but does not imply that there is reproducibility from the analysis data. Level 9 denotes the availability of all the materials (for both analysis and raw data reproducibility) but reproducibility is only achieved from the analysis data (e.g., if there are some issues generating the final analysis data from the raw data). Finally, the top score of 10 denotes the ability to accurately reproduce results starting from the raw data, or **CRR (Computationally Reproducible from Raw Data)**, i.e., unmodified files obtained by the authors from the sources cited in the paper, all the way up to the final display items published in the paper.

Taken together, this means that scores of 5, 9 and 10 in the scale capture either reproducibility from the analysis data (CRA, at scores of 5 or 9) or reproducibility from the raw data (CRR, a score of 10), and they constitute the definition of reproducibility used in the analysis.[9]

Each display item is linked to a scientific claim, which has been defined by the SCORE project (funded by the US Defense Advanced Research Project Agency, DARPA) as "[...] a single major finding from a published study, as well as details of the methods and results that support this finding. A research claim [may not be] equivalent to an entire article. Sometimes the claim as described in the abstract does not exactly match the claim that is tested. In this case, you should consider the research claim to be that which is described in the [results of the paper]". (SCORE-RepliCATS Project)

Said another way, there may not always be a one-to-one correspondence between scientific claims, methods, and results. The SSRP framework allows different claims to be tested using different methodologies, and their results may be presented in one or more distinct display items, such as tables and figures (Figure 1 illustrates the idea that claims, methods, display items and results may not always neatly map one-to-one).

Of course, research claims and results can also have many different structures. Based on two common approaches in empirical economics, SSRP allows for two main high-level categories:

- **Causal claims:** such a claim can be summarized using causal language: "The paper estimates the effect of a variable X on outcome Y for population P, using method M". For example: "This paper (Muralidharan and Prakash 2017) investigates the impact of bicycle provision (X) on secondary school enrollment (Y) among young women in Bihar, India (P), using a Difference-in-Differences approach (M)."
- **Descriptive/predictive claims:** these claims can be summarized using language along the lines of: "The paper estimates the value of a variable Y (estimated or predicted) for population P conditional on X (optional) using method M". For example, "Drawing on a unique Swiss data set (P) and exploiting systematic anomalies in countries' portfolio investment positions

[9] In fact, in the original SSRP scale, scores of 6 or 7 may or may not achieve CRA. Fortunately, information on whether or not CRA is achieved is recorded by the reproducer. In this paper, scores of 6 and 7 that acheive CRA are recoded to a score of 5 for clarity and consistency (and this affects 43 display items across 13 distinct papers). There were also a smaller number of cases where reproducers scored a display item at levels 3, 4 or 8 but also report that it achieved CRA, for instance, after small improvements and/or fixes are made in the code, and these cases are also recoded to a score of 5 or 9 for consistency. These changes affect 20 display items in 10 papers for the move from a score of 3 or 4 to a score of 5, and 18 display items in 13 papers for the move from 8 to 9.

Table 1: Levels of the 10-point SSRP Computational Reproducibility Scale

	Availability of materials, and reproducibility									
	Analysis Code		**Analysis Data**		**CRA**	**Cleaning Code**		**Raw Data**		**CRR**
	P	C	P	C		P	C	P	C	
L1: No materials	–	–	–	–	–	–	–	–	–	–
L2: Only code	✓	✓	–	–	–	–	–	–	–	–
L3: Partial analysis data & code	✓	✓	✓	–	–	–	–	–	–	–
L4: All analysis data & code	✓	✓	✓	✓	–	–	–	–	–	–
L5: Reproducible from analysis	✓	✓	✓	✓	✓	–	–	–	–	–
L6: All cleaning code	✓	✓	✓	✓	(✓)	✓	✓	–	–	–
L7: Some raw data	✓	✓	✓	✓	(✓)	✓	✓	✓	–	–
L8: All raw data	✓	✓	✓	✓	–	✓	✓	✓	✓	–
L9: All raw data + CRA	✓	✓	✓	✓	✓	✓	✓	✓	✓	–
L10: Reproducible from raw data	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

[a] **Computationally Reproducible from Analytic data (CRA):** The output can be reproduced with minimal effort starting from the **analytic** datasets.
[b] **Computationally Reproducible from Raw data (CRR):** The output can be reproduced with minimal effort from the **raw** datasets.
[c] **P** denotes "partial", **C** denotes "complete"

(M), I find that around 8% of the global financial wealth of households is held in tax havens (Y). (Zucman 2013)"

The SSRP approach has several potential advantages. First, the standardization of the reproducibility assessment scale for a specific display item, based on a precisely stated claim in the paper, allows for better comparisons and aggregation, both within and between papers.

The SSRP's standardized and more nuanced approach will, we hope, promote more focused and constructive exchanges regarding assessments of reproducibility. The framework does not typically accommodate simple binary claims such as "paper X does not reproduce at all". Instead, reproductions on the SSRP must clearly point out which claims and tables or figures are being assessed, and up to which level from 1 to 10 they are deemed to be reproducible. To facilitate more constructive and less adversarial exchanges between reproducing authors and original authors, the SSRP guidebook (cited above) contains a chapter with examples for how to conduct more collegial exchanges, including regarding common but potentially thorny issues such as lack of data availability or errors in code.

There remains the issue of how best to aggregate reproducibility scores up from individual display items to a paper-level assessment. As the SSRP is new, norms have not yet evolved among users or in the field of economics as a whole about how best to consider multiple display items relating to a scientific claim, the multiple claims that are made within a paper, and (possibly) multiple distinct

Table 2: Description of the 10-point SSRP Computational Reproducibility Scale

Level	Description
1	No data or code are available. Possible improvements include adding: raw data, analysis data, cleaning code, and analysis code.
2	Code scripts are available (partial or complete), but no data are available. Possible improvements include adding: raw data and analysis data.
3	Analytic data and code are partially available, but raw data and cleaning code are not. Possible improvements include: completing analysis data and/or code, adding raw data, and adding analysis code.
4	All analytic data sets and analysis code are available, but code does not run or produces results different than those in the paper. Possible improvements include: debugging the analysis code or obtaining raw data.
5	Analytic data sets and analysis code are available. They produce the same results as presented in the paper. The reproducibility package may be improved by obtaining the original raw data sets.
6	Cleaning code is partially available, but raw data is not. Possible improvements include: achieving CRA, completing cleaning code and/or raw data.
7	Cleaning code is available and complete, but raw data is not. Possible improvements include: achieving CRA, adding raw data.
8	Cleaning code is available and complete, and raw data is partially available. Possible improvements include: achieving CRA, adding raw data.
9	All the materials (raw data, analytic data, cleaning code, and analysis code) are available. The analysis code produces the same output as presented in the paper. However, the cleaning code does not run or produces different results that those presented in the paper. Possible improvements include: debugging the cleaning code.
10	All the materials are available and produce the same results as presented in the paper with minimal effort, starting from the analytic data or the raw data. Note that Level 10 is aspirational and may be very difficult to attain for most research published today.

reproductions that are carried out for a single published article, i.e., is it the minimum score or maximum score that is most relevant, the average score, or some other statistic? We propose several approaches to dealing with these issues in section 4 below.

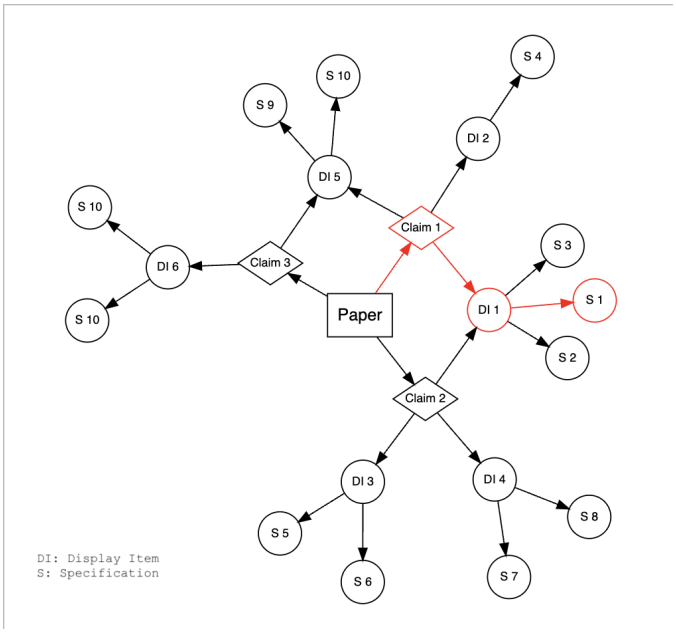
3.2 Beyond Assessments: Improvements and Robustness Checks

While an assessment of computational reproducibility is always required in the SSRP, it is also possible – though optional – to include improvements or robustness checks on the platform, making them publicly available for the research community.

Improvements can come in many different forms, and can be made at the paper level (i.e., a re-organization of the overall file structure or coding files used for the analysis) or could be specific to a display item (i.e., fixing the code to generate the correct standard errors for an estimate). SSRP groups improvements into major categories including (among others): adding data (analysis data or raw data); adding code (for cleaning or analysis); debugging code; or improving instructions on how to access administrative data.

The SSRP also accommodates the inclusion of robustness checks, and these can be recorded in either a more structured or a more open-ended format. The structured approach asks the reproducers to identify the exact lines in a specific file of code that they are changing, followed by a brief summary of how the results change. The more open-ended approach allows for greater flexibility and larger

Figure 1: Stylized Relationship Between a Paper, its Research Claims and its Display Items



departures from the statistical analysis that the original authors had carried out, such that there may not be a specific line of code to exactly build on.

While improvements and robustness checks may end up being among the most important research products posted on the SSRP, for parsimony below we focus mainly on assessments of computational reproducibility of existing studies, in part because of the advantages of the novel quantitative reproducibility scale data.

3.3 A Note on "Abandoned" Reproductions

Reproductions are recorded as "abandoned" when the reproducer attempts to but cannot find any publicly available reproduction materials. In the SSRP, reproducers are advised to leave a short record that (i) no materials exist (or at least were not readily available), (ii) only code is available, or (iii) partial analysis data are available. By the definition provided in the scale, these are level 1, 2, and 3 reproductions, respectively, but they are different from completed reproductions in two important ways.

First, from the point of view of the reproducer it requires relatively little time to submit an abandoned reproduction, just the time searching the web for materials and then a few minutes to record the abandoned reproduction (while a full reproduction exercise can take several hours or even days of work). Second, since abandoned reproductions do not reach the stage of the SSRP where specific claims or display items are assessed, the analysis below treats each abandoned reproduction as if a single claim or display item had been evaluated and assigned a level 1, 2, or 3 score.

The SSRP currently records a total of 175 abandoned reproduction attempts for 173 unique papers (i.e., for a couple of papers, multiple reproducers recorded it as abandoned), all published in economics journals. These come from two sources: 10 reproductions entered by individual SSRP users, and 165 were entered by the SSRP team that had been provided by the Institute for Replication (I4R) through their ongoing activities. Because reproduction exercises on SSRP are often undertaken

for classroom assignments or as ways to deepen engagement with topics of interest by students (see Table A3), reproducers seem likely to self-select papers that already have publicly available reproduction materials. Consequently, reproducers may skip the process of documenting those papers of interest, on which they may have landed initially, for which no code or data are publicly available. Hence there is a likely bias of under-counting the true share of social science papers lacking publicly accessible reproduction materials.

That said, the inclusion of many abandoned reproductions from I4R helps to at least partially address this concern, since the I4R process attempts to more systematically replicate whole sets of recently published articles. Specifically, at the time of writing I4R (which has a related mission to that of SSRP) crowd-sources reproduction/replication attempts by pre-selecting papers from eight leading economic journals from 2022 onwards (and six non-economic social science journals), and has attempted to replicate hundreds of studies. I4R provided a list of papers lacking publicly available replication materials for this analysis, which were then recorded as abandoned reproductions on the SSRP platform.

3.4 Implementation into an Online Platform and Authorship Options

The SSRP framework described above has been developed into a publicly available website at www.socialsciencereproduction.org.[10] Once a reproduction exercise is submitted on the SSRP, it creates a public record. The platform also allows for different degrees (and timing) of availability to the research community. Specifically, reproducers are able to choose among the following three visibility options:

(1) **Fully Public Authorship**: the reproducing authors' names and emails are publicly available, as are the free text (open-ended) answers they provide in their reproduction;

(2) **Class Anonymous Authorship**: reproducers' names are kept private indefinitely (to protect the anonymity of students who perhaps did not even choose which paper to reproduce), as are nearly all of the free text answers they provide in the reproduction, since these are potentially identifying (the one exception being the free text fields that describe the claims being assessed). The main quantitative reproducibility data is publicly available. While the reproducer's identity is not revealed in this case, some key characteristics of the course are made publicly available including the course name, year, the university or institution, the class instructor's name and email address, allowing for a point of contact for scholars interested in following up on the work. Note that students who would like to make their work fully public are still able to do so in SSRP;

(3) **Temporary Anonymous Authorship**: There are sometimes concerns that there could be negative career consequences from carrying out computational reproductions. In light of this, reproducers are allowed to embargo their identity and nearly all free text answers (other than descriptions of claims to be assessed) for up to 4 years after submitting their work on the SSRP. After this embargo period, the record is updated to be fully public authorship and the additional fields and information are made publicly available.

One of the potential uses of the SSRP data is for meta-research of key trends and patterns regarding reproducibility in economics. An initial application of this meta-research agenda is illustrated in the following section, based on the first 487 papers reproduced in the SSRP. Here we explore initial patterns and showcase the types of information and statistics that can be readily produced.

[10]The Appendix contains sample screenshots of an SSRP submission.

4 Descriptive Statistics on SSRP Reproduction Exercises

Since going online in August 2021, 487 reproduction exercises have been completed on the SSRP, corresponding to 380 unique papers. These reproductions assesses a total of 533 claims and 952 display items for their reproducibility (see Table 3). Here we characterize the reports posted to date to better understand the users of the SSRP and explore the state of reproducibility in recent economics research. (While we focus here on papers published in economics journals, results for the smaller number of reproductions from other disciplines are in the appendix.) We first compare the papers reproduced in the SSRP to papers published in the top 100 ranked journals in economics. Then we present the distribution of reproducibility scores recorded in the SSRP, including by aggregating up to the paper level.

Table 3: Number of Items Evaluated on the Social Science Reproduction Platform (SSRP)

Reproduced Item	All Journals	Economics Journals
Paper	380	330
Reproductions	487	411
Claims	738	593
Display Item	952	774
Claim+Display Item	1069	802
Abandoned Reproductions	175	175
Minimum Reproductions	61	38

*As of March 06, 2025

4.1 Papers Reproduced in SSRP Compared to Other Economics Articles

We compare characteristics of papers reproduced on the SSRP relative to a representative (random) sample of papers published in the top 100 journals in Economics (as ranked by IDEAS/RePec as of March 06, 2025) published between 2010 and 2024.[11] To do so, we randomly sampled 15% of papers published in the top 100 Economics journals from Crossref and collected metadata including author names and affiliations, paper citation counts, year of publication, publisher, Digital Object Identifier (DOI), and Journal of Economic Literature (JEL) codes.[12] We also collected the same metadata for all the papers reproduced on the SSRP. All metadata – for both the SSRP papers and the random sample of articles – were collected on the same day, making citation counts and other time-varying data comparable. Paper JEL Codes were first obtained via Crossref API, and when not available there, were sought on various publisher APIs (i.e. Springer, Elsevier) and scraped from publisher websites (i.e. AEA).

Table 4 describes the characteristics of the early papers reproduced on SSRP versus the sample of articles in top 100 economics journals. The SSRP papers have 17% more citations on average than the broader sample, but the median number of citations is similar, indicating that the higher average is likely driven by a small share of highly cited papers on the SSRP. Perhaps related, the SSRP papers are also published in journals with higher impact factors on average, and are more recent than the papers from the top 100 journals (which recall were published between 2010 and

[11] In the representative sample of Economics publications, we were able to collect data on 87 of the top 100 journals, as not all of the top 100 journals seem to be represented on Crossref API. Of the 13 journals missing, only 4 were among the top 50 and 0 among the top 10.

[12] Specifically, we used the python wrapper function crossref_commons.sampling to access the Crossref API.

2024), suggesting that over time the papers on SSRP may be even more heavily cited than the representative sample. Both sets of papers have similar numbers of authors on average.

Table 4: Characterizing Papers Represented on the SSRP

Characteristics of Papers on SSRP vs. in Top 100 Econ Journals

	Papers on SSRP (N=380)					Representative Sample of Econ Papers (N=13342)				
	Mean	Med	Min	Max	%NA	Mean	Med	Min	Max	%NA
Citations	57.9	16.5	0	1781	0	49.6	17	0	5193	0
No. of Authors	2.8	3	1	12	0	2.5	2	1	638	1.7
Yr of Publication	2020.7	2022	1997	2024	0	2017.6	2018	2010	2025	0
Journal Econ Ranking	18.9	18	1	46	17.6	52.2	54	1	100	0
Journal Econ Impact Factor	66.7	55.7	31.8	168.1	17.6	39.4	29.4	19.8	168.1	0
Top Five Econ Journals (%)	40.3					5.1				

Notes: (1) Citation counts as logged by the Crossref's public API as of March 06, 2025, respectively for SSRP and the Universe. (2) Approximately 17.6% of the papers represented on the SSRP are missing ranking and impact factor, as given by the IDEAS/RePec's Top 100 journals in Economics. The majority of missings are due to papers coming from journals' whose primary fields are non-economics disciplines. Papers reproduced on SSRP are from the following disciplines: Economics (87%), Political Science (10%), Medical Science (1%), Demography (1%), etc.

Table 5 lists the ten most common JEL codes for economics sub-fields for both sets of papers. Overall, SSRP papers are drawn more heavily from more empirical applied microeconomic sub-fields, with six of the top seven codes corresponding to Labor Economics (J), Development Economics (O), Health and Education Economics (I), Industrial Organization (L), Public Economics (H), and Agricultural and Environmental Economics. There are far smaller numbers or papers from other subfields (note that the percentages can add up to more than 100% since a single article is typically linked to multiple JEL codes; see Appendix Table A1 for the details).

Table 5: JEL Codes for Economics Papers on SSRP vs. Articles in the Top 100 Economics Journals

	SSRP (All)		SSRP (Ex Abandoned)		Representative Sample	
Top 10 JEL Codes on SSRP	Rank	%	Rank	%	Rank	%
J: Labor and Demographic Economics	1	27.9	2	29.5	5	7.2
O: Development and Growth Economics	2	27	1	33.8	6	6.8
D: Microeconomics	3	25.5	3	25.9	2	12.5
I: Health, Education, and Welfare Economics	4	23.9	4	25.2	11	4.9
L: Industrial Organization	5	12.7	6	9.4	7	6.3
H: Public Economics	6	12.4	5	12.9	10	5.3
Q: Agricultural and Environmental Economics	7	7.9	9	7.9	8	6
K: Law and Economics	8	7	8	7.9	14	1.6
E: Macroeconomics and Monetary Economics	9	6.7	11	6.5	4	7.2
R: Urban and Regional Economics	10	6.7	13	5	13	2.5
	NA	37.88	NA	35.25	NA	55.07

[1] See Table A1 for full rankings.

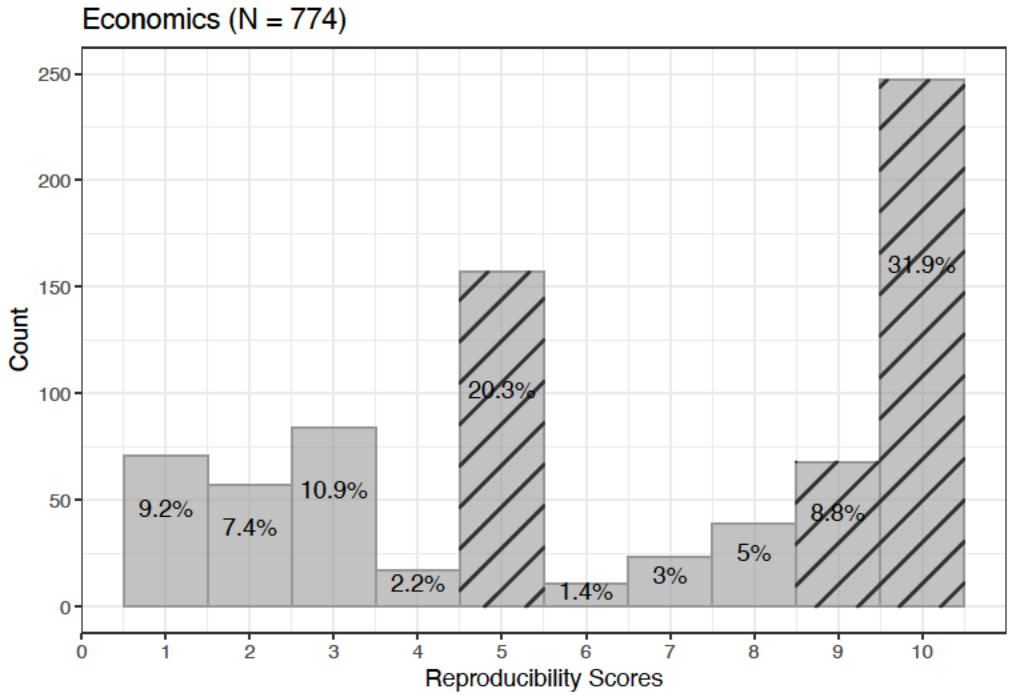
[2] JEL codes may be missing because (i) certain journals or articles do not report them, and (ii) incomplete metadata can hinder retrieval via Crossref and other publisher platforms.

4.2 Reproducibility Scores in the SSRP

We first present the distribution of the scores for each display item (i.e., a table or figure) assessed across the 330 economics papers so far on the SSRP. There are more economics reproductions (411 in all) on the SSRP since some papers have been reproduced more than once. In total, 774 display items have been assigned reproducibility scores on a scale from 1 to 10. Recall that for abandoned reproductions, as explained in Section 3.3, we treat them as if each included a single display item with lower scores of 1, 2 or 3.[13]

The distribution of these scores is presented in Figure 2. This figure shows that there is substantial variation across the display items that have been reproduced to date, with the highest concentration of scores, at 31.9%, at level 10 (meaning full Computational Reproducibility from Raw data, or CRR). An additional 29.1% of display items are rated at level 5 and 9 combined (Computational Reproducibility from Analysis data, or CRA, only, which means the results can be reproduced but the full process of cleaning the data and constructing variables cannot be verified). Thus in all, 61% of display items in the economics papers assessed on SSRP to date are reproducible using at least one of these definitions.

Figure 2: Distribution of Display Item Reproducibility Scores in Economics



[1] For each display item (i.e. tables and figures) assessed, reproducers assign reproducibility scores on a scale of one to ten, as outlined in Tables 1 and 2. The figure is based on papers reproduced on the SSRP through March 06, 2025. Recall that for abandoned reproductions, a single display item is included for each paper.

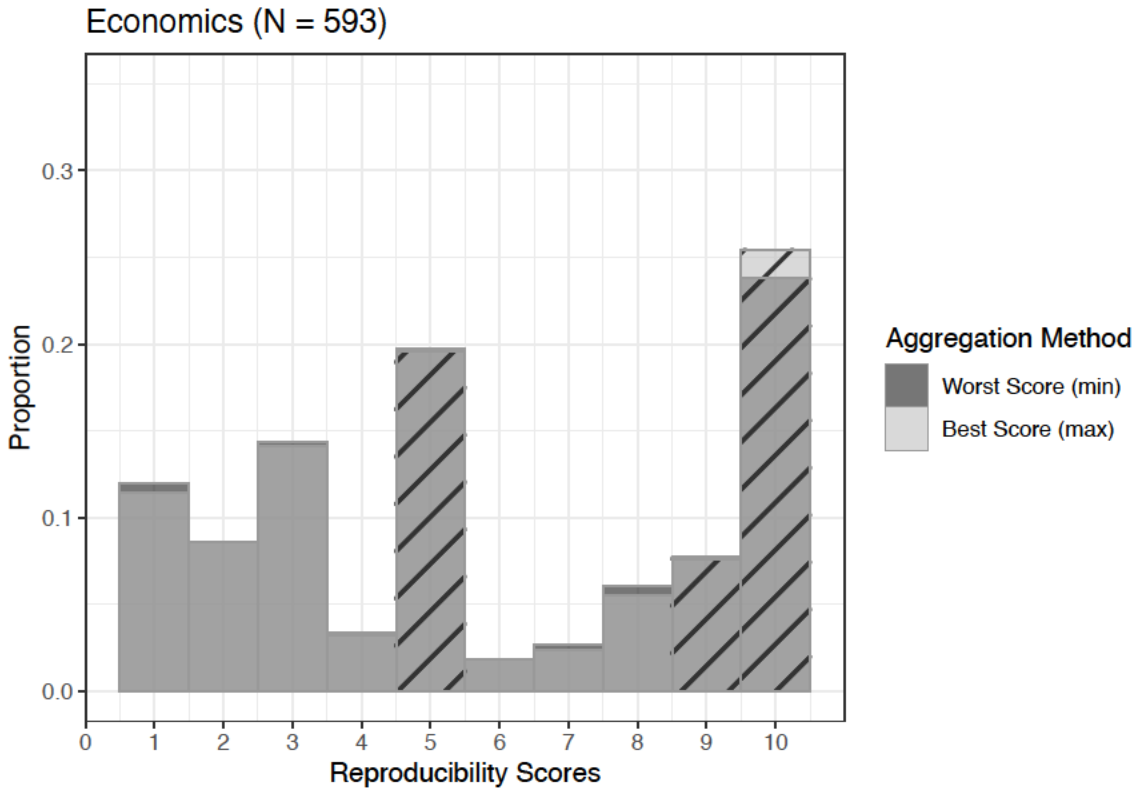
[2] 61% of display items are computationally reproducible from the analysis data (CRA, sum of levels 5, 9, and 10).

Reproducibility scores are next analyzed at the claim level. Recall that a given paper may employ multiple methodologies and display items to support a single claim (or hypothesis), and thus in

[13]Note that this is likely to somewhat underweight the abandoned reproductions in the figure, a concern that we address when we present the paper-level analysis below.

some cases it is necessary to aggregate multiple display item scores (N=774 in the data) up to the claim level (N=593 here). For simplicity, we focus here on two extreme approaches to aggregation across display items: (i) a *pessimistic* approach which takes the minimum score across the display items linked to the same claim, and (ii) an *optimistic* approach that takes the maximum. These are shown visually with different colored shading in Figure 3, and it is visually apparent that the central results do not vary substantially across these two approaches to aggregation, with noticeable spikes again at reproducibility levels 5 and 10, and a range of scores at the other values, including a large cluster at the low scores of 1, 2 and 3.

Figure 3: Distribution of Reproducibility Scores Aggregated to the Claim Level in Economics



[1] A single claim from a paper may be linked to multiple display items. In this figure, the display item level scores are aggregated to the claim level by taking either the minimum or maximum of associated display item scores.
[2] 51% and 53% of claims are computationally reproducible from the analysis data (CRA, sum of levels 5, 9, and 10), aggregating with worst score (min) and best score (max) within each claim, respectively.

We next create a reproduction exercise level score by aggregating scores across multiple claims from the same paper in Figure A.3 (see left graph). (Again recall that some papers have been reproduced multiple times on SSPP, leading to 411 total economics reproduction exercises across 330 distinct articles.) We again either use a pessimistic (min) or optimistic (max) approach across claims. The figure indicates there is considerably more variation in the degree of reproducibility depending on how claims are aggregated within a given reproduction exercise. For instance, the share of reproductions attaining the maximum score of 10 ranges from 17% to 22% depending on whether the more conservative pessimistic definition or the optimistic definition is used. In this case, a natural approach is to focus on the more pessimistic approach, in which case a reproduction exercise would only receive a score of 10 if all claims attain this level.

Finally, we further aggregate up the reproducibility results to the paper level. This differs from the previous analysis since a single paper can be – and often has been – reproduced multiple times by different reproducing authors. For instance, in the SSRP data that we analyze here (with records posted through March 06, 2025), 37 of 330 original papers have been reproduced multiple times, some more than twice.

Figure 4 plots the distribution of scores per paper when taking into account multiple reproductions that have occurred, and using the same approach regarding considering the min or max of scores.[14] Here there is even more variation in the assessments of reproducibility at the paper level depending on whether the min or max approach is used. The proportion of papers scoring a level 10, 9 and 5 are 22%, 5% and 11%, respectively, for a total of 38% when aggregating with the optimistic max approach and 14%, 4% and 12%, respectively, with the pessimistic approach for a total of 30%. Once again, a conservative approach would focus on the minimum, in other words, for a paper to receive a score of 10, all assessments would need to attain this level, and this is only the case in less than a sixth of papers on SSRP (14%). That said, when adding this share to the minimum share at level 5 and 9 – reproducible from analysis data – it appears that roughly a third of papers (30%) are reproducible. Whether or not this is a high or low level is, of course, a matter of perspective, but it is lower than some existing estimates of computational reproducibility in economics. Recall that these figures include the abandoned reproductions, which leads to the large number of papers with low values of 1, 2 or 3.

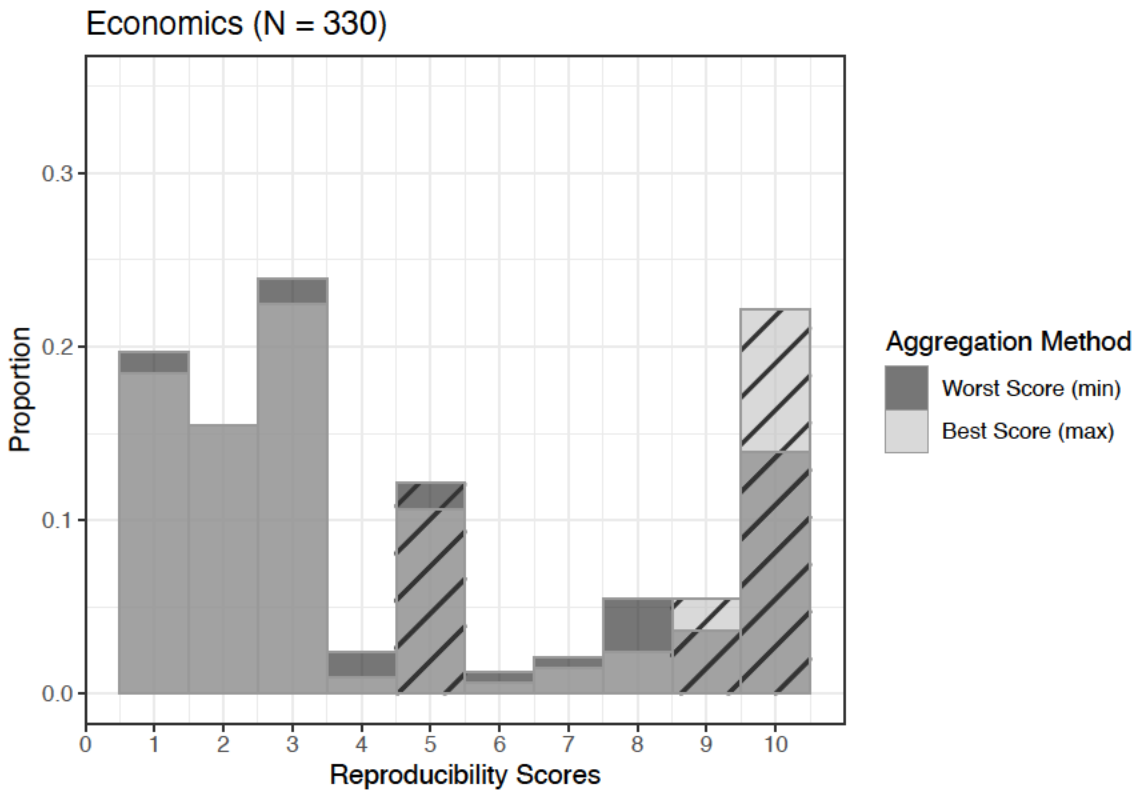
Assessments of the same paper often vary across SSRP reproducers, with some of the variation due to the fact that different reproducers choose distinct display items and/or claims to focus on in their analysis. The ability to compare multiple reproductions of the same paper is a novel contribution of the SSRP platform. As more reproductions per paper become available, it will be possible to compute the statistical reliability of reproducibility scores for a given display item (or paper) with intra-class correlations measures. For now, we are able to compute simpler statistics like the distribution of the maximum difference in scores within each paper: the average absolute difference is 2.92, with nearly half (49%) of papers having an absolute difference of 2 points or less.

4.3 Characteristics of Reproducing Authors

We obtained detailed individual data on the earliest group of reproducing authors on the SSRP, those who had completed reproductions through March 06, 2025, and we characterize them briefly in this section. These reproducers are nearly equally divided between those identifying as women and men, and while the largest share are based at institutions in the US or Canada, there are also sizeable numbers based in Latin America, Asia, and Europe, with smaller numbers in Africa (see Appendix Table A3 for details). In terms of career stage, the majority of reproducers appear to have been students, mainly in graduate school, as expected given that many reproductions are carried out as part of class assignments. Most of the reproducing authors had recorded a single reproduction exercise, but a smaller number had completed multiple ones, with three reproducers having done four or more reproductions to that date.

[14] Similar to above, each statistic is applied consistently from the display item level to the claim level to the reproduction and paper levels, i.e., the min of each or the max of all of them. Figure A.5 shows the same graph excluding abandoned reproductions from I4R.

Figure 4: Distribution of Paper Level Reproducibility Scores in Economics



[1] Multiple reproducers and/or reproduction exercises may evaluate the same paper. In this figure, the reproducibility scores, assessed at the display item level, are aggregated to the paper level, by taking the minimum and maximum from the DI-level reproducibility scores associated with each unique paper. The variation in minimum and maximum reproducibility scores within a paper could come from: (1) variability within a single reproducer/exercise rating distinct display-items with different scores, or (2) variability across multiple reproducers/exercises.
[2] 30% and 38% of reproductions are computationally reproducible from the analysis data (CRA, sum of levels 5, 9, and 10), aggregating with worst score (min) and best score (max) within each paper, respectively.

5 Discussion and Conclusion

The credibility of research findings in economics has come under scrutiny in recent years, with p-hacking and publication bias seen as potential threats to credibility. The adoption of data and code policies by an increasing number of journals now allows researchers to reproduce and replicate published results, increasing the field's ability to more thoroughly appraise, and learn from, published research. The aim of the current study is to describe a platform, the Social Science Reproduction Platform (SSRP), that advances this work by providing a standardized, crowd-sourced, and durable record of reproduction exercises in economics conducted around the world, and to characterize reproducibility given this new source of data.

This initial analysis of the first 487 reproductions in the SSRP has illustrated its uses, its potential to aggregate standardized reproduction assessments, and has provided an initial taste of what a meta-research agenda could look like once the platform becomes even more widely adopted. The analysis shows that there is considerable variation across economics papers in terms of reproducibility. With an aggregation that takes the best score of reproducibility for each paper (across all display

items the reproducer chooses to evaluate in the paper, and across reproductions reproducing the same paper) indicates that roughly 38% of recently published studies in economics appear to meet a reasonable definition of reproducibility (from raw and/or analysis data). When aggregating using the worst score within a paper this quantity falls only slightly to 30% of recent papers. We view this as a disappointingly low rate of reproducibility, although others may have other reactions[15]. An alternative, overly optimistic approach could exclude all abandoned reproductions, as they do not involve a full reproduction exercise. This would change the 38% above to 84%, and 30% to 69%, which is still far from 100%. As the number of reproductions on the platform grows, future work can be used for other types of meta-research documenting patterns in computational reproducibility in economics across sub-fields and over time, or for particular subgroups of scholars or institutions.

The public record of reproductions on the SSRP contain considerable information from reproduction exercises that were previously lost to the research community, and can be used in several ways by the academic community beyond the meta-research illustrated above. First, academic journals, or original authors, could cite or link to reproductions on the SSRP to demonstrate the current reproducibility of a particular paper. Each SSRP reproduction receives a Digital Objective Identifier (DOI) to facilitate such citation. Second, in professional settings (say, for department hiring or promotion decisions), the academic community could use the results of SSRP reproductions as another way to assess the quality of an author's published work. Third, those carrying out the reproductions can use them to demonstrate their in-depth understanding of existing papers and to showcase the improvements and robustness checks they have carried out in relation to the original study, as a valuable new form of scientific knowledge. Fourth, the published SSRP reproductions can then serve as a starting point for further improvements and robustness checks for a given paper.

In all, the SSRP aims to harness the rise in replication and reproduction activities in economics in a way that makes them both comparable and aggregable, with an overall goal of shedding light on how the field can best advance reproducibility, transparency and open science.

[15]Unlike replication exercises where it is not obvious that everything should replicate, we should expect that computational reproducibility is achieved for every paper

References

Andrews, Isaiah and Maximilian Kasy (2019) "Identification of and Correction for Publication Bias," *American Economic Review*, 109 (8), 2766–94.

Ankel-Peters, Jörg, Nathan Fiala, and Florian Neubauer (2023) "Do Economists Replicate?," *Journal of Economic Behavior & Organization*, 212, 219–232.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann (2020) "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics," NBER Working Paper 26993.

Berry, James, Lucas C Coffman, Douglas Hanley, Rania Gihleb, and Alistair J Wilson (2017) "Assessing the Rate of Replication in Economics," *American Economic Review: Papers and Proceedings*, 107 (5), 27–31.

Blanco-Perez, C. and A. Brodeur (2020) "Publication Bias and Editorial Statement on Negative Findings," *Economic Journal*, 130 (629), 1226–1247.

Brandon, Alec and John A List (2015) "Markets for Replication," *Proceedings of the National Academy of Sciences*, 112 (50), 15267–15268.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020) "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110 (11), 3634–60.

Brodeur, Abel, Nikolai Cook, and Carina Neisser (2024a) "P-Hacking, Data Type and Data-Sharing Policy," *Economic Journal*, 134 (659), 985–1018.

Brodeur, Abel, Derek Mikola, Nikolai Cook et al. (2024b) "Mass Reproducibility and Replicability: A New Hope," Institute for Replication Discussion Paper 107.

Buckheit, Jonathan B. and David L. Donoho (1995) "WaveLab and Reproducible Research," in *Wavelets and Statistics*, 55–81: Springer New York, 10.1007/978-1-4612-2544-7_5, ISSN: 0930-0325.

Camerer, Colin F, Anna Dreber, Eskil Forsell et al. (2016) "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351 (6280), 1433–1436.

Camerer, Colin F, Anna Dreber, Felix Holzmeister et al. (2018) "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2 (9), 637–644.

Camerer, Colin F, Anna Dreber, and Magnus Johannesson (2019) "Replication and Other Practices for Improving Scientific Quality in Experimental Economics," *Handbook of Research Methods and Applications in Experimental Economics*.

Chang, Andrew C and Phillip Li (2022) "Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"," *Critical Finance Review*, 11 (1), 185–206.

Cherry Bekaert LLP (2022) "Report of Independent Auditor," 10.1257/aer.112.6.2083.

Coffman, Lucas C and Muriel Niederle (2015) "Pre-Analysis Plans Have Limited Upside, Especially where Replications Are Feasible," *Journal of Economic Perspectives*, 29 (3), 81–98.

Colliard, Jean-Edouard, Christophe Hurlin, and Christophe Perignon (2022) "The Economics of Computational Reproducibility," SSRN: https://ssrn.com/abstract=3418896.

Donoho, David L., Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden (2009) "Reproducible Research in Computational Harmonic Analysis," 11 (1), 8–18, 10.1109/MCSE.2009.15, Conference Name: Computing in Science & Engineering.

Drazen, Allan, Anna Dreber, Erkut Y Ozbay, and Erik Snowberg (2021) "Journal-Based Replication of Experiments: An Application to "Being Chosen to Lead"," *Journal of Public Economics*, 202, 104482.

Gertler, Paul, Sebastian Galiani, and Mauricio Romero (2018) "How to Make Replication the Norm," *Nature*, 554 (7693), 417–9.

Hamermesh, Daniel S (2017) "Replication in Labor Economics: Evidence from Data and what it Suggests," *American Economic Review: Papers and Proceedings*, 107 (5), 37–40.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam et al. (2021) "The Influence of Hidden Researcher Decisions in Applied Microeconomics," *Economic Inquiry*, 59 (3), 944–960.

Ioannidis, John PA, Tom D Stanley, and Hristos Doucouliagos (2017) "The Power of Bias in Economics Research," *Economic Journal*, 127 (605), F236–F265.

Janz, Nicole and Jeremy Freese (2019) "Good and Bad Replications in Political Science: How Replicators and Original Authors (Should) Talk to Each Other," https://www.mzes.uni-mannheim.de/openscience/wp-content/uploads/2019/01/Janz-Freese_-Good-and-Bad-Replications-1.pdf.

McCullough, B.D., K. A. McGeary, and T. D. Harrison (2008) "Do Economics Journal Archives Promote Replicable Research?," *Canadian Journal of Economics*, 41 (4), 1406–1420.

Merton, Robert K (1942) "A Note on Science and Democracy," *Journal of Legal and Political Sociology*, 1, 115, Reprinted as "Science and Democratic Social Structure" in Social Theory and Social Structure ([1949, 1957]1968), pp. 604-615.

Muralidharan, Karthik and Nishith Prakash (2017) "Cycling to School: Increasing Secondary School Enrollment for Girls in India," *American Economic Journal: Applid Economics*, 9 (3), 321–50, https://www.aeaweb.org/articles?id=10.1257/app.20160004.

National Academies of Sciences, Engineering, and Medicine (2019) *Reproducibility and Replicability in Science*: National Academies Press, 10.17226/25303.

Olken, B. A. (2015) "Promises and Perils of Pre-Analysis Plans," *Journal of Economic Perspectives*, 29 (3), 61–80.

Sukhtankar, Sandip (2017) "Replications in Development Economics," *American Economic Review: Papers and Proceedings*, 107 (5), 32–36.

Swanson, Nicholas, Garret Christensen, Rebecca Littman, David Birke, Edward Miguel, Elizabeth Levy Paluck, and Zenan Wang (2020) "Research Transparency Is on the Rise in Economics," in *AEA Papers and Proceedings*, 110, 61–65.

Vilhuber, Lars (2020) "Reproducibility and Replicability in Economics," *Harvard Data Science Review*, 2 (4), 1–39.

——— (2022) "Report by the AEA Data Editor," *AEA Papers and Proceedings*, 112, 813–23, 10.1257/pandp.112.813.

Wood, Benjamin, Rui Müller, and Annette Brown (2018) "Push Button Replication: Is Impact Evaluation Evidence for International Development Verifiable?," https://osf.io/n7a4d/, OSF Preprints.

Zucman, Gabriel (2013) "The Missing Wealth of Nations: Are Europe and the U.S. net Debtors or net Creditors?" *The Quarterly Journal of Economics*, 128 (3), 1321–1364, https://doi.org/10.1093/qje/qjt012.

A Supplemental Appendix

Table A1: Common JEL Codes for Econ Papers on SSRP vs. in Top 100 Econ Journals

JEL Codes	SSRP (All)		SSRP (Ex Abandoned)		Representative Sample	
	Rank	%	Rank	%	Rank	%
J: Labor and Demographic Economics	1	27.9	2	29.5	5	7.2
O: Development and Growth Economics	2	27	1	33.8	6	6.8
D: Microeconomics	3	25.5	3	25.9	2	12.5
I: Health, Education, and Welfare Economics	4	23.9	4	25.2	11	4.9
L: Industrial Organization	5	12.7	6	9.4	7	6.3
H: Public Economics	6	12.4	5	12.9	10	5.3
Q: Agricultural and Environmental Economics	7	7.9	9	7.9	8	6
K: Law and Economics	8	7	8	7.9	14	1.6
E: Macroeconomics and Monetary Economics	9	6.7	11	6.5	4	7.2
R: Urban and Regional Economics	10	6.7	13	5	13	2.5
G: Financial Economics	11	6.4	14	4.3	1	15.7
C: Mathematical and Quantitative Methods	12	6.1	12	5.8	3	11.8
N: Economic History	13	5.5	7	9.4	15	1.2
F: International Economics	14	5.2	18	2.9	9	5.9
Z: Other Special Topics	15	4.8	10	7.2	17	0.8
M: Business, Accounting, and Management	16	3.6	17	3.6	12	2.8
P: Pol. Econ. and Comp. Economic Systems	17	3	15	4.3	16	0.9
A: Gen. Econ & Teaching	18	1.5	16	3.6	19	0.4
B: Economic Thought and Methodology	19	NA	19	NA	18	0.4
Y: Miscellaneous Categories	20	NA	20	NA	20	0.1
	NA	37.88	NA	35.25	NA	55.07

[1] Although papers represented in above table are in economics, JEL codes may still be missing for two reasons: (i) some journals or articles do not present them, and (ii) incomplete publisher metadata sometimes prevents us from retrieving or scraping JEL codes via the Crossref API and other major publisher websites.

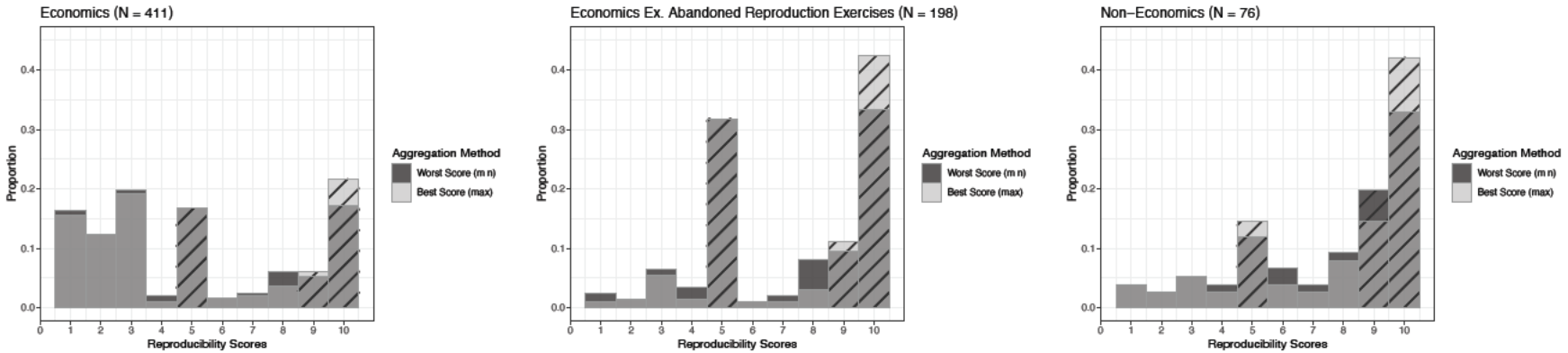
Figure A.1: Dist. of Display Item Reproducibility Scores by Fields

[1] For each display item (i.e. tables and figures) assessed, reproducers assign reproducibility scores on a scale of one to ten, outlined by Table 1 and Table 2.

Figure A.2: Dist. of Reproducibility Scores Aggregated at Claim Level by Fields

[1] A single claim from a paper may be backed by multiple display items. In this figure, the reproducibility scores, assessed at the display item level, are aggregated to claims level, by taking the minimum and maximum from the display item scores associated with each claim. The variation in minimum and maximum reproducibility scores within a claim comes from a single reproducer/exercise rating different display-items that back the same claim with different scores.

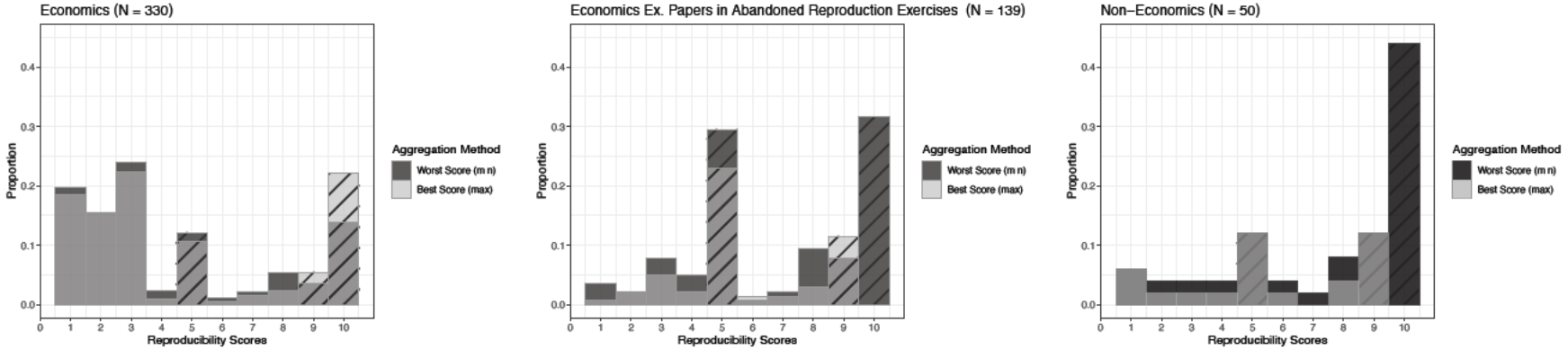
Figure A.3: Dist. of Reproduction Exercise Level Reproducibility Scores by Fields



[1] A Each reproduction exercise may assess multiple display items within a given paper. In this figure, the reproducibility scores, assessed at the display item level, are aggregated to the reproduction exercise level, by taking the minimum and maximum from the DI-level reproducibility scores associated with each exercise. The variation in minimum and maximum reproducibility scores within a reproduction exercise comes from a single reproducer/exercise rating different display-items with different scores.

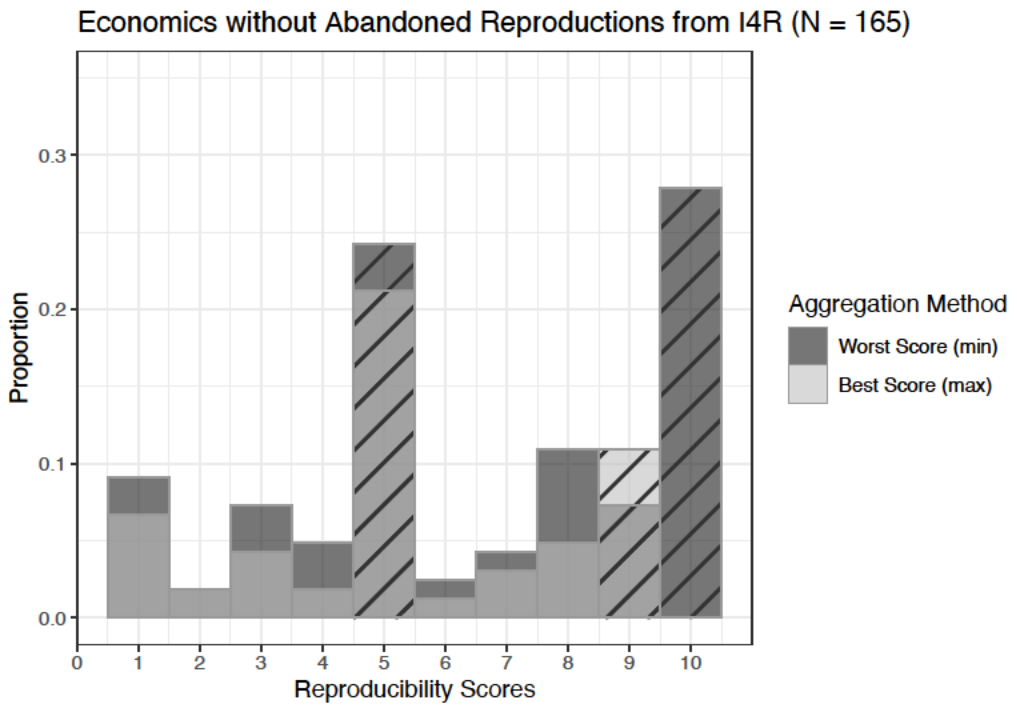
[2] 39% and 45% of reproductions on economics articles (left graph) are computationally reproducible from the analysis data (CRA, sum of level 5, 9, and 10), aggregating with worst score (min) and best score (max) within each reproduction, respectively.

Figure A.4: Dist. of Paper Level Reproducibility Scores by Fields



[1]

Figure A.5: Dist. of Paper Level Reproducibility Scores excl. Abandoned Reproductions from I4R



[1] See section 3.3 for desciption of abandoned reproductions from I4R.

Table A2: Other Assessments from Reproductions

Variable Name / Level	Papers (Max)	Papers (Min)	Display Items
Master File Exists	112 / 380	94 / 380	
Master File One Click	59 / 112	41 / 112	
(All) Output Same	106 / 380		474 / 952
(All) Output Same/Sim	106 / 380		534 / 952

	No. of papers with masterfile that fails to run on one-click	N = 53
Reasons Why Master File Fails	Required To Install Or Update Package / Outdated Command	43
	Path Directory Issues	30
	Software Availability / Discrepancy In Versions	21
	Other Unlisted Error	12
	Files Are Missing	10
	Missing/Undefined Variable	3
	Invalid Syntax	2

* For papers reproduced multiple times by different reproducers, the assessment of whether the master file exists and runs on one-click may differ across reproducers. Hence, when aggregating reproduction level assessment to a paper level, we report two numbers in the numerator, whereby the first number is the upper bound and the number reported in the parenthesis is the lower bound.

Table A3: Reproducer Characteristics (N = 220)

Panel A: Gender	Male	Female					NA
Reproduction	102	91					294
Reproducer	68	64					97
Panel B: Country	U.S.	Canada	Latin America	Asia	Europe	Africa	NA
Reproduction	65	157	44	22	12	4	180
Reproducer	44	97	41	17	2	2	14
Panel C: Education	Undergrad	Grad	Postdoc	Faculty			NA
Reproduction	83	103	12	13			273
Reproducer	67	65	2	3			87
Panel D: Discipline	Econ	Other Soc. Sci.	Math and Comp. Sci.	Natural Sci.	Health Sci.	Info Sci	NA
Reproduction	152	10	4	3	2	10	297
Reproducer	93	8	4	3	2	10	99
Panel E: No. of Reproductions	One	Two	Three	Four	Five or more *		
By Same Authors	171	37	4	4	4		
Panel F: Anonymity	Public	Temporarily Anonymous	Class Anonymous**				
Reproductions	177	16	58				

* The maximum number of reproduction exercises logged by a reproducer or a single SSRP account is 23.

** Reproductions conducted as a part of classroom exercise allows reproducers to remain indefinitely anonymous and instead publicly identifies information about the class and the instructor.

Table A4: Top 100 Econ Journals Scraping Status (Full Universe Before Sampling)

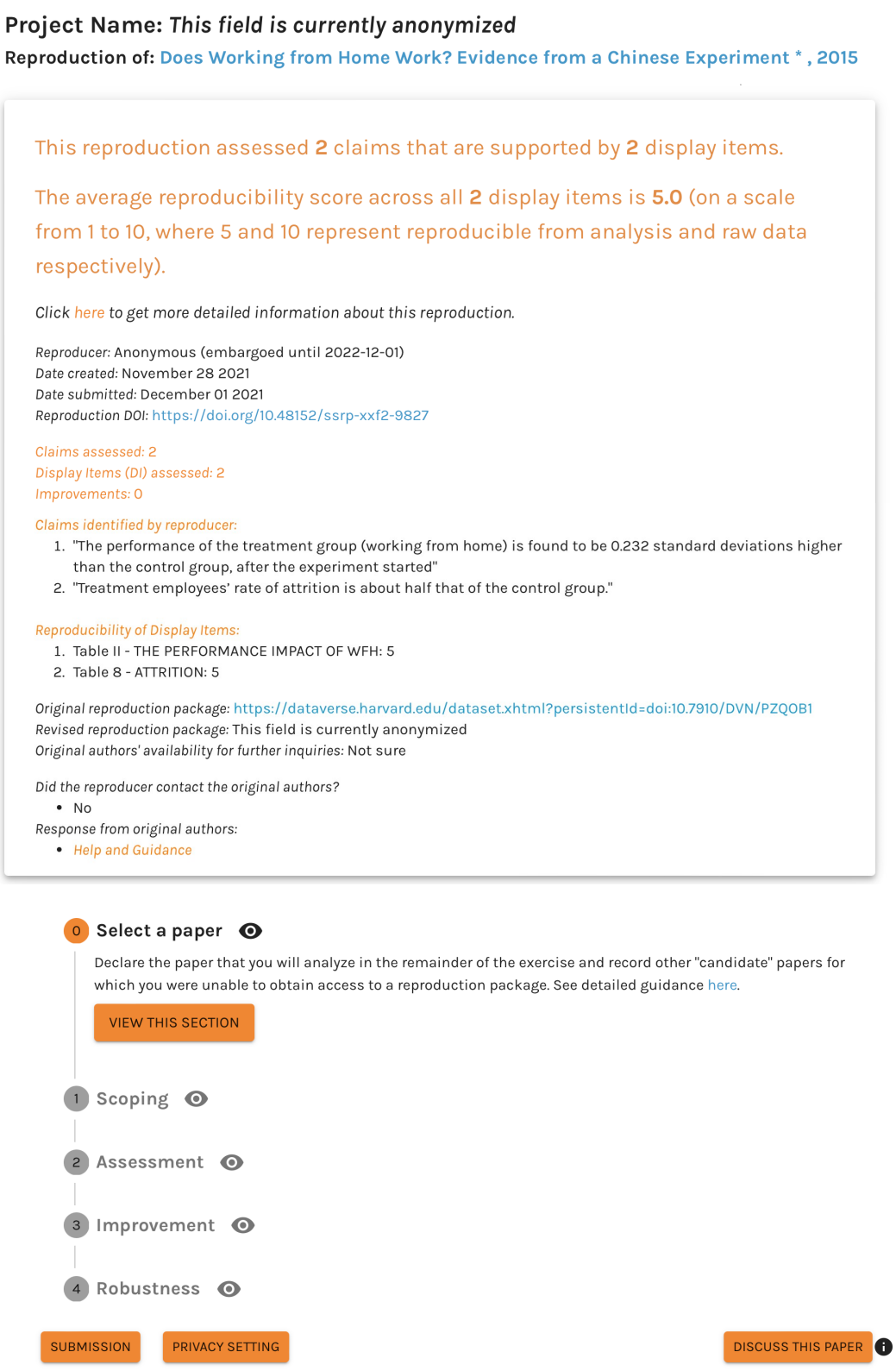
Top 100 Journals - No. of Articles by Year																	
Rank	journal	Scraped	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
1	The Quarterly Journal of Economics	1	44	45	51	41	38	41	44	48	33	38	51	46	44	55	46
2	Econometrica	1	72	52	82	69	64	68	58	65	65	63	80	95	88	72	62
3	Journal of Economic Literature	1	19	24	23	40	19	20	26	33	35	22	24	25	32	55	65
4	Journal of Economic Growth	1	11	17	10	15	9	12	16	12	9	13	11	14	20	19	10
5	Journal of Financial Economics	1	101	134	126	152	101	118	121	123	114	135	151	206	200	88	110
6	Journal of Economic Perspectives	1	54	49	44	49	49	47	49	48	48	48	44	46	44	45	45
7	American Economic Review	1	109	134	136	112	149	126	141	135	114	128	119	114	117	96	111
8	The Review of Financial Studies	1	119	102	91	85	101	83	97	147	129	137	154	141	94	95	90
9	Journal of Finance	1	76	64	64	74	79	79	93	80	76	92	86	81	76	93	101
10	Journal of Political Economy	1	33	36	36	36	38	37	40	70	69	76	117	94	85	93	111
11	Brookings Papers On Economic Activity	1	12	13	12	12	11	10	20	15	14	11	23	10	23	14	6
12	The Review of Economic Studies	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	Proceedings (Federal Reserve Bank of San Francisco)	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	Journal of Accounting And Economics	1	31	33	43	43	33	38	51	43	43	44	38	52	49	47	70
15	Journal of Monetary Economics	1	76	47	60	62	76	85	68	50	59	63	100	138	74	70	106
16	American Economic Journal: Macroeconomics	1	38	32	31	29	31	35	30	32	32	40	38	44	53	55	50
17	Journal of The European Economic Association	1	113	51	43	59	69	40	30	66	61	82	69	57	71	77	58
18	American Economic Journal: Applied Economics	1	44	41	44	41	40	35	37	42	46	46	43	37	57	56	64
19	Rand Journal of Economics	1	34	32	32	32	35	36	39	43	37	35	45	36	27	23	28
20	Journal of Labor Economics	1	29	25	27	33	29	40	52	40	42	55	36	54	45	49	102
21	Foundations And Trends(R) In Econometrics	1	3	2	0	3	3	1	1	1	1	1	1	2	3	2	3
22	Economic Policy	1	18	15	16	16	21	12	16	17	17	15	19	14	65	48	46
23	The Review of Economics And Statistics	1	78	101	80	119	71	78	67	73	70	67	64	66	91	211	266
24	Journal of Econometrics	1	142	148	175	100	153	199	153	127	126	160	150	160	137	276	266
25	Strategic Management Journal	1	66	83	86	170	133	108	139	91	127	105	113	97	97	80	98
26	The Economic Journal	1	74	84	70	80	101	90	72	195	83	77	141	103	94	113	121
27	Annual Review of Economics	1	24	25	23	18	37	24	22	22	22	36	31	24	33	28	23
28	Journal of Applied Econometrics	1	86	56	68	53	54	78	63	44	64	56	51	75	53	66	73
29	Journal of Business Venturing	1	44	43	41	47	48	49	36	39	44	50	56	67	54	35	50
30	Proceedings (Federal Reserve Bank of Cleveland)	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
31	Journal of Business & Economic Statistics	1	42	50	51	46	69	39	51	118	68	74	79	109	110	72	90
32	Journal of International Economics	1	71	69	90	99	82	90	103	94	91	78	93	93	94	101	122
33	Journal of Financial Intermediation	1	27	27	28	34	23	27	25	27	36	23	28	34	39	33	26
34	Journal of Public Economics	1	94	139	92	106	137	117	85	121	143	98	134	156	130	169	153
35	Journal of International Business Studies	1	66	54	38	59	72	52	56	70	75	88	83	94	98	83	85
36	The World Bank Research Observer	1	14	7	7	11	17	10	6	5	6	12	7	8	8	11	8
37	American Economic Journal: Economic Policy	1	35	34	41	48	45	46	39	52	48	54	51	57	64	64	64
38	Organization Science	1	73	102	101	95	92	99	85	61	64	66	73	69	111	116	137
39	The World Bank Economic Review	1	18	36	21	41	14	62	68	12	25	72	31	27	27	46	61
40	Post-Print Ceb	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41	Journal of Development Economics	1	73	80	105	111	122	70	81	71	121	89	131	124	118	151	151
42	Research Policy	1	117	117	136	140	139	140	158	135	157	197	145	155	185	174	138
43	Journal of Human Resources	1	75	56	72	66	65	65	43	40	46	60	67	74	74	65	48
44	Journal of Economic Surveys	1	48	30	36	48	41	39	40	40	46	37	43	69	44	49	63
45	Foundations And Trends(R) In Finance	1	2	3	1	2	5	3	0	3	0	0	2	2	1	2	2
46	Journal of Law And Economics	1	32	52	32	33	43	32	29	24	24	24	26	28	45	28	27
47	Experimental Economics	1	35	43	34	37	36	45	30	49	38	42	49	47	49	32	30
48	Journal of Environmental Economics And Management	1	40	57	59	70	54	47	46	78	128	93	70	105	88	109	128
49	Journal of Money, Credit And Banking	1	77	102	88	80	90	91	60	58	118	69	51	124	84	99	95
50	Journal of Risk And Uncertainty	1	26	25	20	24	24	24	20	21	24	24	24	25	30	19	27
51	Journal of Accounting Research	1	39	18	37	34	31	32	24	21	34	26	37	58	40	45	48
52	Proceedings (Board of Governors of The Federal Reserve System)	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table A4: Top 100 Econ Journals Scraping Status (Full Universe Before Sampling) *(continued)*

Rank	journal	Scraped	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
53	Quarterly Review	1	2	1	2	0	1	0	1	0	0	1	3	0	0	2	4
54	International Economic Review	1	48	51	55	52	51	48	56	48	114	32	76	59	55	61	61
55	Review of Economic Dynamics	1	43	41	31	43	44	51	51	51	52	59	68	54	56	102	22
56	Journal of Consumer Research	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
57	Imf Economic Review	1	15	24	22	23	22	39	29	19	28	26	22	27	41	35	43
58	The Journal of Business	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
59	Journal of Urban Economics	1	55	35	47	46	53	41	48	55	51	40	40	60	56	59	63
60	Econometric Theory	1	54	48	50	44	80	36	47	38	37	24	50	52	51	25	33
61	Journal of Economic Theory	1	113	111	108	109	129	168	113	104	132	113	115	154	153	131	114
62	European Economic Review	1	63	72	108	104	118	110	157	142	128	144	195	201	158	215	181
63	Journal of Banking & Finance	1	249	253	254	396	298	297	171	170	235	193	214	186	243	229	162
64	Oxford Bulletin of Economics And Statistics	1	50	65	53	45	31	34	27	70	53	58	55	61	48	52	53
65	International Journal of Central Banking	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
66	The Econometrics Journal	1	18	28	35	22	22	18	15	23	16	28	33	34	32	29	25
67	The Journal of Law, Economics, And Organization	1	19	25	49	26	20	27	20	15	25	22	26	41	28	33	29
68	Journal of Financial And Quantitative Analysis	1	67	55	44	61	53	50	68	90	170	81	90	66	139	133	38
69	Review of Environmental Economics And Policy	1	22	19	18	19	19	17	20	25	22	20	20	21	19	21	20
70	Mathematical Finance	1	82	20	53	41	30	20	20	20	30	31	39	41	25	53	22
71	Bis Quarterly Review	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
72	Journal of Global Economic Analysis	1	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0
73	Journal of Health Economics	1	75	103	70	110	82	98	82	82	97	79	85	116	88	85	68
74	Marketing Science	1	74	74	56	57	48	52	52	53	51	50	60	54	57	60	102
75	Energy Economics	1	170	145	272	262	235	260	283	357	340	381	374	546	566	651	738
76	Journal of Economic Geography	1	52	51	50	57	54	51	46	48	67	43	34	49	37	39	49
77	Journal of International Money And Finance	1	96	94	105	146	127	116	120	121	99	139	102	149	161	149	171
78	Annual Review of Financial Economics	1	14	21	19	12	18	21	12	14	21	11	15	19	22	31	18
79	Journal of Financial Markets	1	20	23	19	27	46	21	28	29	32	29	32	36	56	62	36
80	Proceedings (Federal Reserve Bank of Philadelphia)	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
81	Review of Finance	1	25	37	39	71	52	68	70	65	40	23	41	39	71	44	45
82	Economic Policy Review	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
83	Journal of Corporate Finance	1	47	96	78	75	121	108	121	157	129	103	236	281	140	139	126
84	Economía Journal	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
85	Journal of Industrial Economics	1	42	29	26	30	27	27	31	32	25	16	26	29	31	65	22
86	Foundations And Trends(R) In Microeconomics	1	8	2	3	3	5	0	0	1	3	0	0	1	1	0	0
87	American Journal of Agricultural Economics Appendices	1	142	187	119	128	118	95	103	94	105	71	137	94	78	73	57
88	Financial Management	1	60	37	43	31	30	40	33	27	46	38	27	37	18	22	39
89	Journal of Empirical Finance	1	62	63	51	63	79	65	99	61	76	61	61	74	61	102	79
90	Journal of The Royal Statistical Society Series B	1	36	35	43	52	48	38	68	35	37	39	47	58	79	161	121
91	Accounting, Organizations And Society	1	51	43	34	38	40	46	41	37	41	37	43	35	32	34	49
92	Entrepreneurship Theory And Practice	1	57	65	58	76	64	54	69	63	73	43	65	59	66	33	44
93	Oecd Journal: Economic Studies	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
94	Scandinavian Journal of Economics	1	38	51	49	54	52	36	31	42	45	95	29	30	36	40	33
95	Management Science	1	163	150	159	180	180	186	207	243	326	284	295	380	438	424	679
96	Oxford Review of Economic Policy	1	35	31	40	33	38	22	31	50	31	35	62	46	52	58	60
97	Journal of Population Economics	1	47	47	57	38	36	39	42	54	40	32	51	63	51	45	79
98	Labour Economics	1	98	95	85	84	78	62	76	85	106	81	116	84	153	113	135
99	Small Business Economics	1	53	102	53	68	82	79	120	150	147	164	133	140	133	142	128
100	Journal of Financial Stability	1	24	20	25	64	61	74	83	84	85	52	47	87	79	62	120

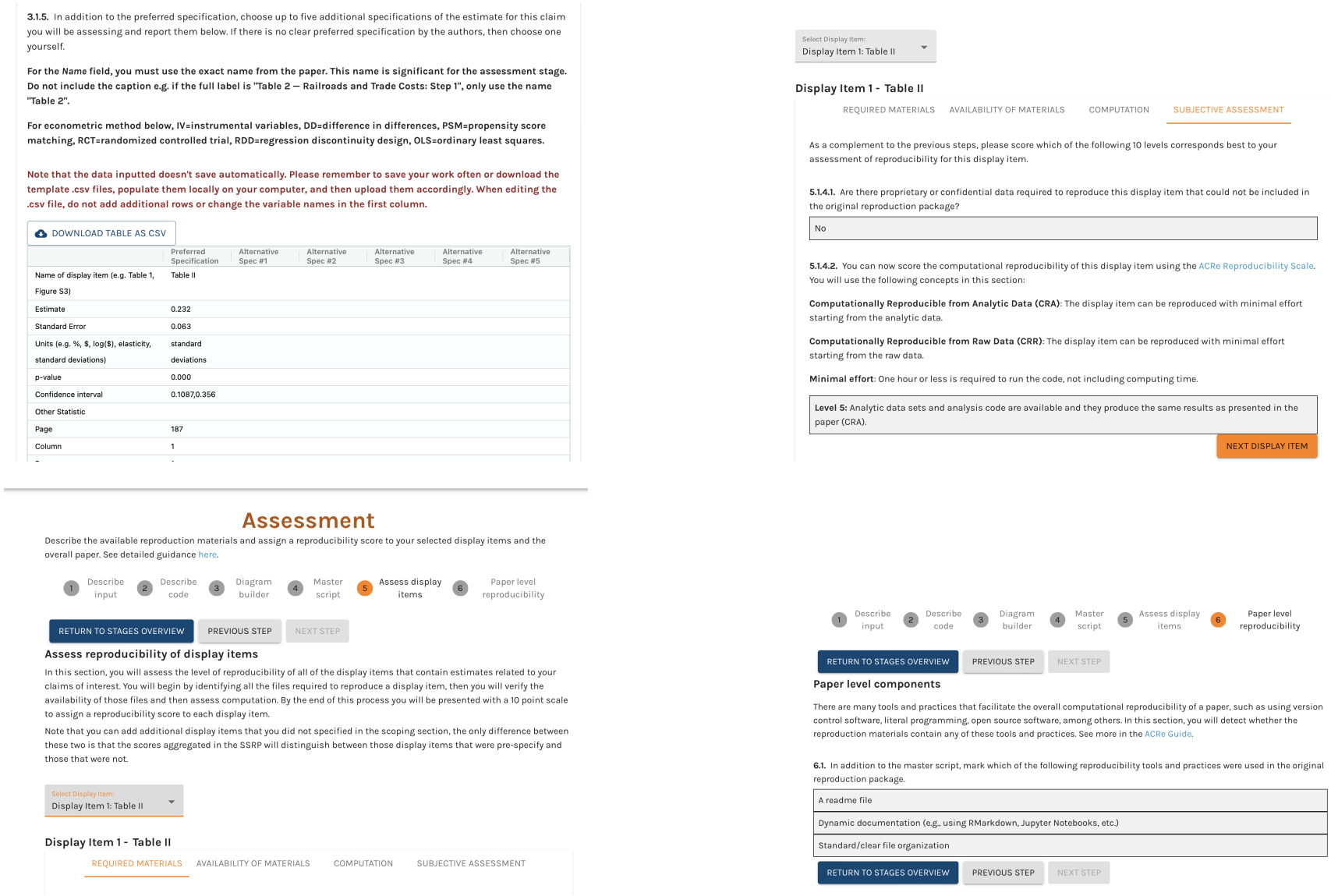
[1] Note: Number of articles are counted by unique DOIs by journal-year as identified by Crossref after basic filtering based on article title; non-research items such as data editor notes and audit reports are excluded. These counts may be approximate.

Figure A.6: Reproduction Summary



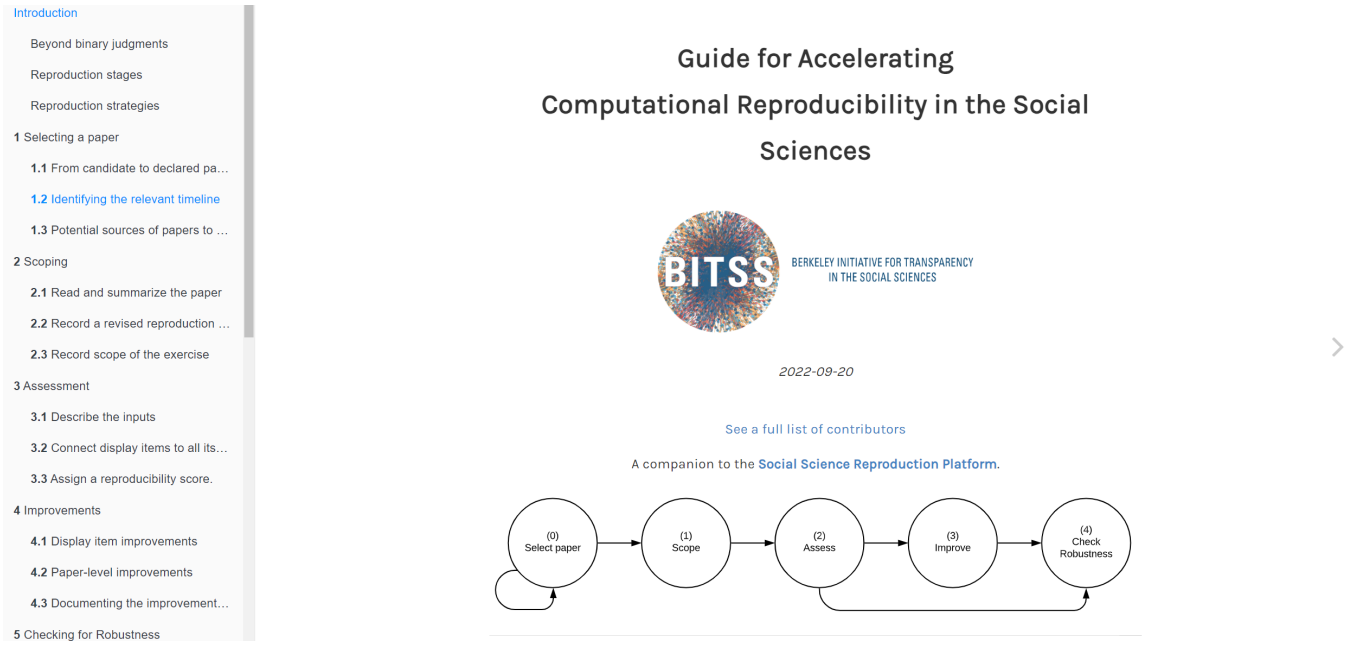
Top figure is an example of an anonymously submitted reproduction exercise of Bloom et al. (2015). The summary page on SSRP identifies the claims and display items reproduced and the average reproducibility score assigned to the display items. Bottom figure demonstrates different sections of the submitted reproduction exercises and a link to the paper's discussion forum.

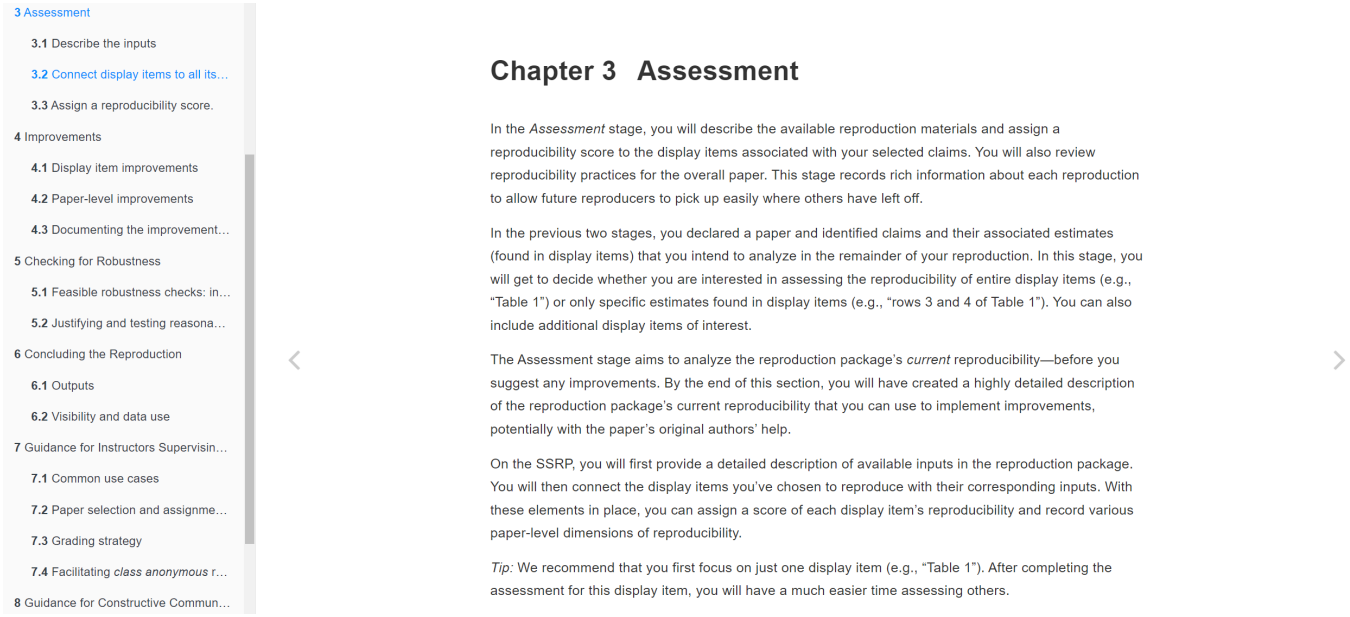
Figure A.7: Detailed Scoping and Assessment



Above figure showcases how a claim and associated display items are identified and assessed in several steps. Top-left panel identifies econometric methods and estimates associated with a claim. Bottom-left panel begins the assessment of a claim at the display-item level. Top-right panel demonstrates scoring of each display-item. Bottom-right panel collects information on other reproducibility-related characteristics at the paper-level.

Figure A.8: Detailed Scoping and Assessment





Accompanying ACRE guide provides the standardized conceptual framework for reproduction exercises. See Section 3.1 for details.