



Teaching for Large-Scale Reproducibility Verification

Lars Vilhuber, Hyuk Harry Son, Meredith Welch, David N. Wasser & Michael Darisse

To cite this article: Lars Vilhuber, Hyuk Harry Son, Meredith Welch, David N. Wasser & Michael Darisse (2022) Teaching for Large-Scale Reproducibility Verification, Journal of Statistics and Data Science Education, 30:3, 274-281, DOI: [10.1080/26939169.2022.2074582](https://doi.org/10.1080/26939169.2022.2074582)

To link to this article: <https://doi.org/10.1080/26939169.2022.2074582>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC



[View supplementary material](#)



Published online: 23 Jun 2022.



[Submit your article to this journal](#)



Article views: 514



[View related articles](#)



Citing articles: 1 [View citing articles](#)

Teaching for Large-Scale Reproducibility Verification

Lars Vilhuber , Hyuk Harry Son, Meredith Welch, David N. Wasser, and Michael Darisse

Cornell University, Ithaca, NY

ABSTRACT

We describe a unique environment in which undergraduate students from various STEM and social science disciplines are trained in data provenance and reproducible methods, and then apply that knowledge to real, conditionally accepted manuscripts and associated replication packages. We describe in detail the recruitment, training, and regular activities. While the activity is not part of a regular curriculum, the skills and knowledge taught through explicit training of reproducible methods and principles, and reinforced through repeated application in a real-life workflow, contribute to the education of these undergraduate students, and prepare them for post-graduation jobs and further studies. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received September 2021
Accepted April 2022

KEYWORDS

Economics; Reproducibility;
Undergraduate training

1. Introduction

The purpose of scientific publishing is the dissemination of robust research findings, exposing them to the scrutiny of peers. Key to this endeavor is documenting the provenance of those findings. Recent years have seen significant concerns expressed about the robustness of scientific results, commonly referred to as the “replication crisis” (King 1995; Hamermesh 2007; Gall et al. 2017; Fanelli 2018). Various facets of the “crisis” have been explored (for just some of these, see Hamermesh 2007; Olken 2015; Camerer et al. 2016; Stodden et al. 2018). Various approaches and solutions have been called for and proposed by the National Academies (National Academies of Sciences, Engineering, and Medicine 2019), committees of National Science Foundation (Bollen et al. 2015), and many scientists have called for greater transparency of research practices, and more assurance that published research is reproducible (Bell and Miller 2013; Stodden et al. 2016; Höffler 2017b; Clemens 2017; Coffman et al. 2017). Learned societies have a role to play within this discussion, as do journals.¹

We should note here that the terms “reproducible” and “replicable” are not well-defined. Throughout this article, we will use them as defined in Bollen et al. (2015) and National Academies of Sciences, Engineering, and Medicine (2019): (computational) reproducibility is “obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis” (National Academies of Sciences, Engineering, and Medicine 2019, p. 36). Replicability is achieved in this context through a relaxation of certain of the constraints implicit in the definition of reproducibility, for instance by collecting new data, implementing different methods or code, and then


“obtaining consistent results across studies aimed at answering the same scientific question [...] [obtaining] consistent results given the level of uncertainty inherent in the system under study” (*ibidem*). These definitions are broadly accepted in the social science and statistics community nowadays, but other communities may actually use these terms somewhat differently (e.g., Heroux 2015). We refer to “replication packages” as the collection of materials provided by authors to enable *others* to replicate the results, but which should be “reproducible” themselves.

For empirical articles, the foundations on which they reside (data and its analysis) are external to the article and often to the journal they are published in. The data posting policies of many societies and journals, including the American Economic Association’s pre-2019 policy (Bernanke 2004), were and are intended to create a minimal framework from which to replicate empirical findings. Historically, they have often failed. In several studies (Camerer et al. 2016; Chang and Li 2017; Höffler 2017a), at least half of the replication packages associated with surveyed manuscripts failed to (fully) reproduce the results in the manuscript when re-run.

Increasingly, societies and journals have therefore, switched to verifying and monitoring these policies (Jacoby et al. 2017; Vilhuber 2019; Editors 2021). The American Economic Association (AEA), the largest association of professional and academic economists in the world, with over 20,000 members located in 148 countries, has been at the forefront of such policies. It publishes eight journals, including one of the top five journals in the discipline, in addition to several well-respected field journals. Concerns about the reliability and robustness of economic

CONTACT Lars Vilhuber  lars.vilhuber@cornell.edu  Cornell University, Ithaca, NY

¹For a collection of articles describing reproducibility in various disciplines, including economics (Vilhuber 2020), as well as an overview of the NASEM report cited above (Fineberg et al. 2020), see articles in the *Harvard Data Science Review’s* Fall 2020 issue.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/UJSE.

© 2022 The Author(s). Published with license by Taylor and Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

research have circulated in the AEA's membership for more than 30 years (Dewald et al. 1986; McCullough and Vinod 1999). The policy to require that articles provide copies of their replication materials, first implemented in 2004 (Bernanke 2004), was highly innovative at the time, but reflective of the membership's requests. Nevertheless, these early efforts improved the availability, but not necessarily the reproducibility of these replication packages. In 2018, the Association appointed one of the authors of this article (Vilhuber) as the inaugural Data Editor (Duflo and Hoynes 2018). The Data Editor, in turn, started verifying, prior to final acceptance of an article, the computational reproducibility of the results displayed in the manuscript. The AEA's endeavor is the largest in scale amongst the journals and societies conducting such verifications, having verified more than 1,000 articles since initiating such verifications two years ago.

The verification of replication packages, which checks not only the computational reproducibility of the provided materials, but also verifies the documented provenance and completeness of such materials, is not a magical solution that will solve the "reproducibility crisis." Replication packages may be reproducible, but wrong (see the recent discussion surrounding Simonsohn et al. 2021). They do, however, reduce the cost for the scientific community to more easily find and assess such issues (finding the issue documented in Simonsohn et al. (2021) would have been much harder without a complete replication package) and can find other issues much earlier. For instance, the recent retraction in the *Journal of Finance* (2021) would have been detected prior to publication, not two years after publication. In the case of restricted-access data, prepublication verification, when possible, may sometimes be the only opportunity to conduct such checks. Whether conducting reproducibility checks prior to publication is the most efficient or conducive exercise remains an open issue, including within the AEA's discussions on this topic. In this article, we describe the AEA's activities and how they contribute to this discussion.

The AEA began conducting comprehensive prepublication reproducibility verification for conditionally-accepted manuscripts at its eight journals starting in 2019. These checks are conducted by the Labor Dynamics Institute (LDI) Replication Lab, which was set up by the current AEA Data Editor (Vilhuber), and which we describe in more detail in the next section. The Lab hires and trains undergraduate students who are primarily responsible for performing the required checks. The Lab's work with students is not integrated into any curriculum. Nevertheless, we will argue that students acquire some of the key data and computational skills described in National Academies of Sciences (2018). These abilities are the result of both the Lab's training and the observation of numerous completed but imperfect research projects. We also argue that students gain "hands-on" experience with some of the key dimensions of data-oriented science. The typical student will work on dozens of replication packages during their time at the Lab, with increasing autonomy along the way. These packages are of varying levels of complexity and completeness, and the students are required to assess their compliance with evolving and multifaceted standards. This combination of taught and experiential learning provides the students with a strong foundation in data and computational management.

The goal of this article is to describe the setting, the selection process for students, the actual workflow, and sketch out some of the observed outcomes. We hope to show that, while the setting may currently be fairly unique in its scale and position within the academic publication cycle, it is feasible to implement in a broader setting, and can meaningfully contribute to students' data science education.

2. Methods

2.1. Setting

The AEA's Data and Code Availability Policy (DCAP) states that *[i]t is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is non-exclusive to the authors* (American Economic Association 2019, 2020). To achieve the goal of improved (pervasive) computational reproducibility, the authors of conditionally accepted manuscripts are required to submit a replication package, consisting of all code used to process and analyze data, any data not available in an existing trusted repository, and a "README" describing data provenance and processing instructions.² Each replication package is assessed in terms of data provenance, clarity of the description, and computational reproducibility. These checks are conducted by the LDI Replication Lab (henceforth "the Lab").

The Lab was created by one of the authors of this article and AEA Data Editor (Vilhuber), based on earlier work starting in 2014, described in Kingi et al. (2018). The Lab hires and trains undergraduate students who are primarily responsible for performing the required checks. These students are supervised by the Data Editor, the assistant Data Editor, and a graduate student. In a given calendar year, a total of about 50 undergraduate students work in the Lab with approximately 15–20 actively working on replications in a given week. Including the pilot phase, the Lab has hired over 100 Cornell University undergraduate research assistants (RAs) to complete reproducibility verification checks.

While not a primary objective of the Lab's work for the AEA, the reliance on undergraduate students is intentional. The project provides undergraduate students with a unique opportunity to gain insights into the research and publication practices of hundreds of economists every year. Students are trained on reproducibility techniques, best practices, and communication skills, addressing several of the key dimensions highlighted in the National Academies of Sciences (2018) report, namely "data acumen" and knowledge of some "computational foundations," "data management and curation," as well as "workflow and reproducibility" skills. Upon graduation, these students in turn take those skills into nonacademic workplaces or graduate studies, thus, seeding the next generation's improved practices. The Lab, therefore, attempts to target both ends of the academic

²The current requirements for authors are described in much detail at <https://aeadataeditor.github.io/aea-de-guidance/>, but have varied over time. Confidential data are not part of the replication package, but must be described in the README, and are regularly made available to the Data Editor privately. As of February 2022, the README *should* conform to Vilhuber et al. (2020), but this requirement, too, has varied over time.

publication pipeline: the “outlet,” by improving replication packages of conditionally accepted publications, and the “inlet,” by improving skills and attitudes of students at an early stage in their career. This article will focus on the latter.

The type of activity conducted at the Lab (systematic computational reproducibility checks within a publication workflow) is relatively novel. In particular, we are not aware of other institutions conducting such activities with undergraduate students. Other institutions that conduct reproducibility checks, such as the Odum Institute (Christian et al. 2018) and *casca*d (Pérignon et al. 2019) primarily conduct such work with graduate students and professionals.

2.2. Recruiting

The Lab relies on students having some prior experience with statistical software, so as to be productive in short order. Despite the growth of open source software in the last decade, the software used in top economics journals is still overwhelmingly proprietary software (Stata, Matlab; see Figure 2 in Vilhuber 2020). Furthermore, most usage by economists of said software is still strongly desktop-oriented. This is reflected in the replication packages received, and determines the required skill set of the students. However, the operating paradigms for most of these software packages remain quite similar. We have found that experience in any one of these packages is sufficient to allow students to follow instructions and conduct reproducibility verifications. Through prior experience, we have also found that knowledge of software more frequently used in computer science or engineering (for instance, Java, or C++) was not so useful, driving our current requirement of experience with what we know to be similar computing paradigms. Of note, while we require some exposure to statistical software commonly used in the social sciences, students are not expected to master it, nor to be proficient programmers.

These requirements are described in job postings, circulated among various undergraduate student experience coordinators on campus, and published to the campus-wide employment opportunity website and the LDI website.³ The posted job description provides applicants with information on requirements, wage rate, and maximum hours. As of January 2022, undergraduate RAs in the Lab are hired at the starting end of the Level II (out of four) classification wage rate, which, as of December 31, 2021, is \$13.45 per hour.⁴ Hours are limited to 10 hr per week, which is less than the maximum number of hours allowed by Cornell University policy.⁵ Students are recruited from across campus, but most applicants are from the social sciences. Students apply with a cover letter and a CV, and are selected based primarily on observed and self-declared experience with one of the common statistical software packages used in the social sciences. The field they are majoring in plays no role, but is correlated with experience in these statistical software packages. Since most of the manuscripts stem from

economics, most applicants have an interest, and often a major, in economics. No explicit years of study requirement is targeted. In practice, we have hired second through fourth-year students, but not first-year students. In the most recent two rounds, we had 48 applications, and selected 21 for training. Of the 48 applications, 21 mentioned economics as one of their majors, 7 statistics, 9 computer or information science, and 21 none of those. Of the 21 students selected for training, 16 had an economics major, 5 had a statistics major, and 4 had a computer or information science major, with 4 having declared none of those majors.⁶

2.3. Training

Although we announce the minimum requirements upon recruitment, the skill levels of the trainees are often heterogeneous, which could raise problems in conducting reproducibility assessments. Therefore, we train applicants before making final hiring decisions. The training is not remunerated, but is also not meant as a test. Rather, our intention with the training is to upskill all applicants. In practice, we retain over 90% of trained applicants for at least one semester, and nearly all attrition in the past has been voluntary.

We currently provide training three times a year: before the fall and spring semesters, for students joining the Lab in that semester, as well as at the end of the spring semester for students joining the Lab as a summer job. During this initial training, we provide instructions on all essential skills and knowledge necessary for the tasks. The base training includes an overview lecture on the context of reproducibility concerns in economics, provides knowledge on reproducible practices, data provenance and data citations (Data Citation Synthesis Group and Martone 2014), includes a presentation on the Vilhuber et al. (2020) README, basic instructions on command line and version control systems, and a detailed walkthrough of the assessment process, including how to prepare the final reproducibility report sent back to the authors.⁷ It also reinforces computational skills, but does not train students on basic computational skills (since that was part of the selection criterion). Many of these topics are new to undergraduate social science students.

Depending on circumstances, in particular during the unusual 2020–2021 period, base training has been conducted as an intense 8-h training day, as a sequence of 2–4 hr training sessions spread over multiple days on Zoom, or even as an 8-h long virtual training session. The intense basic training is followed by a sequence of targeted test cases, interspersed with additional short lectures, reinforcing and deepening certain aspects (data provenance, debugging), as well as helping students acclimate to the Lab’s task scheduling, reporting and workflow systems. Again, in adapting to changing circumstances over the last several semesters, these test cases and additional lectures have been concentrated into the three days immediately following the

⁶Statistics generated from internal records as of September 15, 2021.

⁷The agenda for the training can be viewed at <https://labordynamicsinstitute.github.io/replicability-training/>, most of the slides are available at <https://labordynamicsinstitute.github.io/replicability-training-presentation>, and the textual content of training, which the presentation loosely follows, is available at <https://labordynamicsinstitute.github.io/replicability-training-curriculum/>.

³The posting as it appears on the LDI website as of February 2022 can be viewed at <https://perma.cc/G969-2HT4>.

⁴The full pay scale as it appeared to students in February 2022 can be viewed at <https://perma.cc/PTB5-98ZW>.

⁵Students can and do work more hours when classes are not in session—January and the summer months.

base training, or stretched out over the next three weeks, or only the following week. Students work on each test case on their own, following detailed step-by-step instructions, then interact with more experienced peers (undergraduate students already employed by the lab), and finally submit the report for each test case (as described in the Workflow) to the senior instructors and Lab leaders. Each case is discussed as a group before the next test case is presented and assigned, and students receive both generic feedback (commonly made errors or omissions) and individualized feedback.⁸

The first test case uses the dominant software package Stata, and introduces students to several small impediments to reproducibility. The case is a simple fake article with one table. The notion of “add-on packages” (libraries, etc., that need to be installed) is introduced. Stata, R, and many other statistical software rely on such packages, but they need to be specified for a replication package to be considered complete, since replicators may not have these installed. Authors of replication packages often neglect to identify such packages, and if not available, code will fail. The first test case uses one such package, but does not specify that it is needed. Students learn to recognize the type of error this generates, how to solve the problem, and how to document both the existence of the problem and the solution. Second, students are introduced to the idea of publicly accessible data that cannot be provided as part of a replication package. In this case, the package uses a dataset made available by ICPSR to researchers at member institutions. The terms of use specify that the data should not be redistributed. Students therefore, need to download the data themselves. They learn to document any requirements (such as registration requirements, costs, application procedures, etc.), and to evaluate whether the description of the access conditions in a README is complete and sufficient. Finally, this test case is the first exposure to the structured replication template,⁹ the use of git and Markdown, and how to navigate the workflow process.

The second test case uses the second-most frequent software in economics, MATLAB. Students are again provided with an article, except this time, it is a real article, with all its complexity and idiosyncracies. Students are now faced with identifying whether all code is provided for the tables, figures, and numbers in the article. It is a surprisingly frequent problem with replication packages that some significant part of the code is not provided. Sometimes, the missing code is used for appendices, sometimes it is data cleaning code, and sometimes, it is a key part of the overall replication package. In this case, both data cleaning code and code for some manuscript figures are missing. As for data, all the data appear to be provided, and students are challenged to identify whether data provenance is sufficiently documented. Numerically, it is a “small data” case, so it is actually possible to compare the source data with the provided data. Finally, for many students, the use of MATLAB pushes most of them out of their comfort zone, and part of the training is to guide them to the similarities in user interfaces of statistical software, rather than focusing on the differences in the programming languages. The second test case is also the

second time that the workflow is navigated, and the template used, instilling familiarity with the tools that will be used in the Lab.

The third test case introduces students to more data provenance issues, while making it harder to verify that all the data are present. The case is a real article that uses confidential data that cannot be shared. Thus, students are challenged to identify whether all the data appear to be documented, by reading not just the README, but also the data section in the article. The article relies on Scandinavian register data and includes a thorough discussion of this data (a luxury not always present in other articles that students will later encounter). Students will assess whether data are properly cited, and write the report without ever running any code.

Both the second and third test cases rely on articles that were subsequently published.¹⁰ These cases were chosen because they give students practice with common issues they will face, while not being too computationally difficult or time consuming. The second test case emphasizes some of the technical skills needed to evaluate reproducibility (running code in an unfamiliar program, identifying parts of code in relation to tables/figures, etc), while the third test case helps students learn and practice how to identify data sources and verify packages for completeness without being able to run code. Students are shown how the final, published article and package differ from the earlier versions that they were provided with, and can see the impact that the reports can have on the clarity and reproducibility of scholarly publications.

As noted, each of the training test cases is accompanied by an initial presentation of the topic, significant independent work that relies on prepared documentation, a mentoring session with an experienced undergraduate replicator, and a debriefing session in which students provide their first impressions, and instructors provide both specific and generalized feedback.

At the end of the initial training, students join the regular Lab meetings, and are assigned real cases. However, we do generally take care to assign them cases that we believe to be easier to complete, slowly ramping up the complexity. We can do this because we have overlapping cohorts of students, with varying levels of experience, present in the Lab.

2.4. Activities

We now describe the generic workflow for the verification activities conducted in the Lab. A simplified workflow is shown in Figure 1, a more detailed workflow with specific instructions for students is publicly available and regularly updated.¹¹ The Lab receives a request to verify a replication package, generating a case number. Such replication packages generically consist of (or should consist of) computer code, possibly but rarely software, data that can be redistributed, and a document, generally referred to as the “README,” describing the provenance of all data, including data not made available as part of the package,

⁸The individualized feedback is generally prepared by one of the graduate assistants.

⁹<https://github.com/AEADDataEditor/replication-template>.

¹⁰The students work on publicly available preprint versions of the articles, which do not include any of the recommendations made by the Data Editor and incorporated by the authors into the final version of record.

¹¹<https://labordynamicsinstitute.github.io/replicability-training-curriculum/aea-jira-workflow-a-guide.html>.

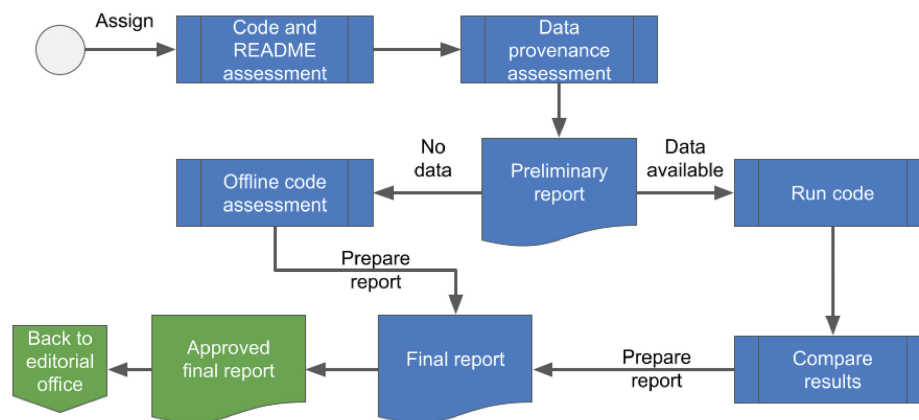


Figure 1. Simplified workflow for verification activities in the LDI Lab.

and the process to reproduce the results in the paper. Materials are usually deposited by authors at the AEA Data and Code Repository,¹² though deposits made at other trusted repositories are also accepted (but are rare). The associated manuscript is also made available to the Lab.

The case is assigned to the next available student. At this stage, no allowance is made for student skills or experience. The first part consists of an assessment of the README, the manuscript, and of the package contents.¹³ Students assess whether data provenance is completely described, whether the manuscript and/or the README contain data citations (Data Citation Synthesis Group and Martone 2014), and whether there appears to be a complete archive of the program code (consistent with the README). Data provenance is assessed and verified, even when data are provided. When data are not provided, data provenance is even more important, as it must completely describe how the student can obtain the data. This can range from simple click-through downloads to complex application processes. The computer code and the README are assessed in terms of the software requirements. Finally, the computational resources are summarized, if reported in the README or deduced from other descriptions.

This information is summarized in a preliminary report. One of the longest tasks is the data assessment task, since some manuscripts combine several dozen distinct data sources, and the quality of the description still ranges from excellent to poor. Once the preliminary report is finalized, it may be discussed with senior Lab members (team leads, graduate students, Data Editor): Does the assigned student have the necessary skills? Does the Lab have access to the software? Does the Lab have access to the computing resources (specific OS, high-performance computing)? Does the Lab have access to the data, or can it obtain access to the data without an unreasonable delay? Do the computations run in a reasonable amount of time, given all the resources available? Based on these assessments, the case may be reassigned to a student with the requisite experience,

to a cooperating third party with access to data or computing resources, or, most frequently, remains with the original student.

The student assigned to conduct the computational reproducibility check then downloads any necessary data, and follows instructions in the README to run the programs provided. This can be as simple as running a single “main” controller program, or as complex as running and baby-sitting dozens of programs, as well as conducting certain activities by hand. While much software can be automated, it is frequently automatable tasks are conducted manually. Most often, this involves parsing computer output to generate tables (despite perfectly good code to do so programmatically) and generating maps using GIS software (despite perfectly good automated software to do so). It is the student’s job to reproduce the manuscript’s results following the authors’ instructions as closely as possible, though this has some limits, in particular for very inefficient code or instructions.

While all of the above steps appear to be very systematic, there are many subjective assessments. When are the manual steps too onerous? When is a deviation from stated computer runtimes unusual or too long, and action must be taken? The students are queried frequently about progress, and are required to discuss their active cases at twice-weekly meetings. In those meetings, all cases are discussed openly, and decisions are made. Those decisions do not necessarily come from the Data Editor, although he is present at most of these meetings. Students are encouraged to propose reasonable solutions, based on their experience, and increasingly make decisions autonomously. Input to such decisions may also come from team leads, or from graduate students. Students can also use a mailing list for peer-to-peer conversations and problems, and are encouraged to contact the student pre-approvers and the senior Lab members (Data Editor, assistant data editor, and graduate student) with more specific problems, which may be resolved in one-on-one meetings.

Once the results have (or have not) been reproduced, the student completes a report. While based off of a template, the report has many free-form and narrative components, describing the steps the replicator undertook to achieve full computational reproducibility. Students are trained and mentored in how to convey steps in a concise but complete way, allowing the reader to understand fully what was done, and why it may not

¹²<https://www.openicpsr.org/openicpsr/aea>.

¹³We note that the AEA states that authors *should* use the standardized README template (Vilhuber et al. 2020). Many of the issues that the Lab encounters could be addressed by correct use of the README template.

have worked. Students have access to a bank of frequently used “canned responses”¹⁴ that provide constructive feedback and a checklist of requests. They are also mentored in providing an objective, positive, and dispassionate tone, as they are communicating with much more senior members of the profession.

The report is reviewed by more experienced students (“pre-approvers”),¹⁵ and then reviewed again by the Data Editor before being sent to authors.¹⁶ Students get feedback on how they performed on each case, and when improvements need to be made to the report or to the methods employed, they are provided both directly (privately) to the student, as well as in the form of generalized and anonymized counsel in the group meetings.

Most reports require that authors improve their package. In some cases, in particular when computational reproducibility was not achieved for key tables and figures, the replication package will go through another cycle of review after resubmission by the authors. When possible, such resubmissions are assessed by the same replicator. This is done both because it is efficient (the replicator will already have downloaded all the necessary data, often many gigabytes, into their personal replication workspace) and because it provides validating feedback to the replicator, showing them that authors are (generally) quite responsive to constructive comments. Students then write an abbreviated revision report, which goes through the same review cycle as the full initial report.

The cycle described above gets repeated for every new case assigned to the Lab. The typical case will be in the student’s hands for about 10–14 days, though they may sometimes work on two or more cases simultaneously, as they wait for code to run or data access requests to be authorized, and some cases may take significantly longer.

To give context, we refer the reader to the AEA Data Editor’s annual reports (Vilhuber 2019; Vilhuber, Turitto, and Welch 2020; Vilhuber 2021, 2022). In the most recent year, the LDI Lab received 529 requests for 415 manuscripts (Vilhuber 2022). The vast majority of manuscripts go through a single round of reviews, typically with minor changes requested (Vilhuber 2022, Tables 2 and 3). The median time to full acceptance is between four and six weeks (reported in various annual reports by the AEA journals).¹⁷

3. Discussion

The recruitment, training, and regular activities have been refined in an iterative process since early 2018. After each training, we have reviewed the effectiveness of our methods, surveyed trainees about their perception of the training, and incorporated improvements into the next round. For instance,

¹⁴The “canned responses” can be found at <https://perma.cc/U8MR-WEEZ>. The latest version of the canned responses is available in the students’ cloned template every time they initiate a new verification.

¹⁵We do not discuss the supplementary training for pre-approvers here.

¹⁶A sample report, suitably anonymized, can be viewed at <https://perma.cc/Z577-EEHG> and is included in Appendix A, *supplementary materials*.

¹⁷As an aside, when replication packages are not unconditionally accepted, changes that are requested almost always include both corrections to data provenance descriptions (incomplete access description, absent data citations) and minor, fixable corrections to code (directories not created, requirements not fully described).

the peer-driven tutorial as part of the test cases was based on trainee feedback, and has proven both popular and effective.

3.1. Student Development Opportunities within the Lab

As students accumulate cases, we have observed fairly rapid gains in maturity and autonomy. The best students may be promoted to team leads, where they run the shorter check-in meetings mostly autonomously, are expected to provide a first level of support to their team members, and may be asked to be “pre-approvers.” Students contribute regularly to overall improvements in processes and procedures, and are encouraged to contribute to a Wiki, with the intent of providing a knowledge base that is driven by the students for the students. In certain circumstances, students may provide short (verbal or written) tutorials or presentations at group meetings when they have solved a particularly thorny problem, for which the solution is of potential utility to all the students.

3.2. Student-Faculty Interaction at the Lab

Two types of student-faculty interaction occur during the Lab activities in which the students participate. First, students meet twice a week with the faculty supervisor (the Data Editor), to discuss individual cases. These meetings are group meetings, and take as long as necessary to guide the student onto the right path. Where appropriate, one-on-one meetings with the faculty supervisor or graduate students are also scheduled, to solve thornier problems.

The second type of interaction is more indirect. All reports written by students, once approved, are read by manuscript authors, and may be read by journal editors. Students also read the authors’ responses to any requested changes. Authors and editors are most often academic faculty. Students do not, however, interact directly with authors and editors - in fact, this is explicitly forbidden by the rules of behavior of the Lab. Thus, this indirect interaction with academic and private-sector faculty is always mediated via structured written communication.

3.3. Relationship to other Curricular Activities

The activities described here are not offered as a regular for-credit course. Rather, they are offered as an on-campus, research-oriented job. As such, they do not neatly fit into a curricular development or plan. However, our impression is that students gain valuable experience through participation in the Lab’s activities. This (admittedly subjective and potentially biased) opinion relies on a few observations. For one, almost all students stay with the Lab until graduation, and sometimes even beyond graduation. They also tend to continue at the Lab over the summer, despite having other summer jobs or internships (the hours are adjusted to accommodate such constraints). Such implicit “job satisfaction” indicates that students believe that the experience is valuable, despite having access to higher-paying jobs during summers and post-graduation.

Second, while most students come to us with some prior exposure to programming and data analysis, they largely need training on both the conceptual and technical skills necessary to

assess reproducibility. This suggests that they are not currently acquiring these skills elsewhere in their coursework. While the various curricula on campus do offer some of these elements explicitly, and other courses likely embed these techniques within their more discipline-specific topics, there does seem to be a need to more generally expose undergraduate students to these techniques. The Lab's training, while not designed to provide training in such techniques, seems to complement other curricular offerings. For example, while many students have read academic papers in other courses, they rarely have the skill set upon joining the lab to identify and summarize the data sources used in the article. This skill is something that we train students on in order to evaluate the existence or completeness of data citations.

3.4. Longer Term Outcomes

We have not previously conducted a formal evaluation of the takeaways and experiences undergraduates have had after working for the Lab, and we do not systematically conduct exit interviews with undergraduates at the end of their job tenure. We have, however, conducted informal conversations with the goal of a formative evaluation of initial training and later activities. A few graduates have reported that they have “learned a lot about reproducibility” in ways that “really helped me as a research assistant” (2021 economics graduate working for a large national economic consulting firm), and that they received “overwhelmingly positive feedback on my documentation method in code reviews, which is all thanks to my time with LDI” (2020 sociology graduate working for a nonprofit research organization). A more formal survey of former Lab members is currently in the field, and we plan to report on outcomes in the future.

One of the difficulties of empirically measuring the effect of the explicit initial training and the implicit on-the-job training provided is the noisiness of most empirical measures. Students are not required to actively program, but they will learn programming techniques. No student will have been exposed to data citation principles prior to joining the Lab, and yet all will have learned and applied such principles by the end of the initial training, and refined it over time. The students' efficiency at conducting computational reproducibility invariably will increase, but allocation of papers is not random, and the difficulty of papers varies so widely that any objective measure of time or effort will likely be too noisy to be useful. We continue considering ways to measure this in the future.

3.5. Generalization of Lab Activities

When initially planning how to translate the AEA's verification activities into a feasible operation, the reliance on undergraduate students was intentional, for two reasons. For one, we believed that with proper training, undergraduates would provide a more cost effective verification of the basic computational reproducibility of the packages we received than with a rotating cadre of graduate students. This has largely been born out, though we do not provide evidence here on the financial underpinnings of the operation. More importantly, however, we intentionally forced ourselves to develop a training program for those undergraduates, believing that it would have

utility for other universities, disciplines, journals, and in other circumstances.

If an instructor wanted to directly integrate these activities into a curriculum, it is likely best to implement them as a formal course. This course should include the training that research assistants currently receive as well as training for how to actively create reproducible packages on their own. The initial training alone accounts for about 14 hours of classroom instruction, plus significant “homework” time. With a straightforward and pedagogically valuable expansion of some of the themes that get short shrift in the current training because they are taught and learned in the first “real cases” (git, version control, objective communication skills), a course based on these materials would easily cover 21 hours of classroom training. Test cases could be expanded to have higher and more varied computational requirements, and can easily be based on real replication packages, suitably modified to highlight specific learning objectives. Student assessments could be based on recognition of required components of reproducible packages, followed by written reports on actual replication packages, and concluding with the students creating their own replication package, based on either a research project from another course or a novel one as part of this course. If using a research project from another course, then this replication course could be paired with, for example, an upper level course with a term paper component.

While it may seem attractive to embed some of the minor activities into an existing course, our experience with the particular student population from which we recruited leads us to conclude that the necessary startup training is intensive enough to require a standalone course. We note, however, that almost none of our recruits have taken more explicitly data-science-oriented courses, as far as we can tell. Our suggestions are thus likely not representative when more technically advanced students are included in the population of interest.

We do, however, believe that what we observe is not specific to economics, and could be easily implemented in other (related) disciplines. Discussions with colleagues in sociology or policy analysis suggests that the basic context and student skill set may be quite similar there. Amongst our recruits have been engineering, bio-statistics, and sociology students, and they have performed just as well as the students with an economics major.

One final thought, though. Implementing the activities outlined here as part of a course is unlikely to then also meet the needs of a journal for reliable and timely verification service. Most journals would have verification needs at monthly or even weekly frequencies throughout the year, including when classes are not in session. A course would not even produce a continuous stream of verification reports throughout the semester. A course might, however, serve as a feeder to a campus-wide verification service, similar to statistical consulting services with student workers.

Supplementary Materials

A sample report, referenced in Footnote 16, is included in Appendix A and also available at <https://perma.cc/Z577-EEHG>. Additional supporting materials are the “canned responses” (see Footnote 14), available at <https://perma.cc/U8MR-WEEZ>, the job posting (see Footnote 3),

available at <https://perma.cc/G969-2HT4>, and the student pay scale as of February 2022 (see Footnote 4), available at <https://perma.cc/PTB5-98ZW>. Additional material supporting the active training as described here can be found at <https://labordynamicsinstitute.github.io/replicability-training/>, <https://labordynamicsinstitute.github.io/replicability-training-curriculum-labordynamicsinstitute.github.io/replicability-training-presentation>, and <https://github.com/AEADDataEditor/replication-template>. All materials are available under CC-BY or similar licenses.

Acknowledgments

We thank Hautahi Kingi, Sylvérie Herbert, and Flavio Stanchi, who contributed to the earliest pilots of this effort. We thank the students who have worked for the Lab, and who have helped improve both the reproducibility of economics articles, as well as the training and workflow for later participants.

Disclosure Statement

LV is the Data Editor of the American Economic Association, a paid position. All activities described herein are funded by the American Economic Association. HHS, MW, and DNW were graduate student assistants at various points during the time period described in the article, and MD is the assistant data editor, all hired by LV and paid through the AEA's contract with Cornell University.

ORCID

Lars Vilhuber  <http://orcid.org/0000-0001-5733-8932>

References

- American Economic Association (2019), "Data and Code Availability Policy." Available at <https://www.aeaweb.org/journals/data/data-code-policy>
- American Economic Association (2020), "Data and Code Availability Policy," *AEA Papers and Proceedings*, 110, 776–78.
- Bell, M., and Miller, N. (2013), "How to Persuade Journals to Accept Your Replication Paper." Available at <https://politicalsciencereplication.wordpress.com/2013/09/11/guest-blog-how-to-persuade-journals-to-accept-yourreplication-paper/>
- Bernanke, B. S. (2004), "Editorial Statement," *The American Economic Review*, 94, 404–404.
- Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., and Olds, J. L. (2015), "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. Available at https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016), "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436.
- Chang, A. C., and Li, P. (2017), "A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time," *American Economic Review*, 107, 60–64.
- Christian, T.-M., Lafferty-Hess, S., Jacoby, W., and Carsey, T. (2018), "Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals," *International Journal of Digital Curation*, 13, 114–124.
- Clemens, M. A. (2017), "The Meaning of Failed Replications: A Review and Proposal," *Journal of Economic Surveys*, 31, 326–342.
- Coffman, L. C., Niederle, M., and Wilson, A. J. (2017), "A Proposal to Organize and Promote Replications," *American Economic Review*, 107, 41–45.
- Data Citation Synthesis Group and Martone, M. (2014), "Joint Declaration of Data Citation Principles," Technical report, Force11. Available at <https://doi.org/10.25490/a97f-egy>
- Dewald, W. G., Thursby, J. G., and Anderson, R. G. (1986), "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," *The American Economic Review*, 76, 587–603.
- Duflo, E., and Hoynes, H. (2018), "Report of the Search Committee to Appoint a Data Editor for the AEA," *AEA Papers and Proceedings*, 108, 745.
- Editors (2021), "Supporting Computational Reproducibility through Code Review," *Nature Human Behaviour*, 5, 965–966.
- Fanelli, D. (2018), "Opinion: Is Science Really Facing a Reproducibility Crisis, and do We Need it to?" *Proceedings of the National Academy of Sciences*, 115, 2628–2631.
- Fineberg, H., Stodden, V., and Meng, X.-L. (2020), "Highlights of the US National Academies Report on 'Reproducibility and Replicability in Science,'" *Harvard Data Science Review*, 2.
- Gall, T., Ioannidis, J. P. A., and Maniatis, Z. (2017), "The Credibility Crisis in Research: Can Economics Tools help?" *PLOS Biology*, 15, e2001846.
- Hamermesh, D. S. (2007), "Viewpoint: Replication in Economics," *Canadian Journal of Economics*, 40, 715–733.
- Heroux, M. A. (2015), "Editorial: ACM TOMS Replicated Computational Results Initiative," *ACM Transactions on Mathematical Software*, 41, 1–5.
- Höfler, J. H. (2017a), "Replication and Economics Journal Policies," *American Economic Review*, 107, 52–55.
- (2017b), "ReplicationWiki: Improving Transparency in Social Sciences Research," *D-Lib Magazine*, 23.
- Jacoby, W. G., Lafferty-Hess, S., and Christian, T.-M. (2017), "Should Journals Be Responsible for Reproducibility?" Inside Higher Ed.
- Journal of Finance (2021), "Retracted: Risk Management in Financial Institutions," *The Journal of Finance*, 76, 2709.
- King, G. (1995), "Replication, Replication," *PS, Political Science & Politics*, 28, 443–499.
- Kingi, H., Stanchi, F., Vilhuber, L., and Herbert, S. (2018), "The Reproducibility of Economics Research: A Case Study," Presentation, Berkeley, CA. Available at <https://osf.io/srg57/>
- McCullough, B. D., and Vinod, H. D. (1999), "The Numerical Reliability of Econometric Software," *Journal of Economic Literature*, 37, 633–665.
- National Academies of Sciences, Engineering, and Medicine (2018), *Data Science for Undergraduates: Opportunities and Options*, Washington, DC: The National Academic Press.
- (2019), *Reproducibility and Replicability in Science*, Washington, DC: National Academies Press.
- Olken, B. A. (2015), "Promises and Perils of Pre-analysis Plans," *Journal of Economic Perspectives*, 29, 61–80.
- Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R., and Debonnel, E. (2019), "Certify Reproducibility with Confidential Data," *Science*, 365, 127–128.
- Simonsohn, U., Nelson, L., Simmons, J., and Anonymous. (2021), "[98] Evidence of Fraud in an Influential Field Experiment About Dishonesty."
- Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., and Taufer, M. (2016), "Enhancing Reproducibility for Computational Methods," *Science*, 354, 1240–1241.
- Stodden, V., Seiler, J., and Ma, Z. (2018), "An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility," *Proceedings of the National Academy of Sciences*, 115, 2584–2589.
- Vilhuber, L. (2021), "Report by the AEA Data Editor," *AEA Papers and Proceedings*, 111, 808–817.
- (2022), "Report by the AEA Data Editor," *AEA Papers and Proceedings*, 112, 813–823.
- (2019), "Report by the AEA Data Editor," *AEA Papers and Proceedings*, 109, 718–729.
- (2020), "Reproducibility and Replicability in Economics," *Harvard Data Science Review*, 2. DOI: 10.1162/99608f92.4f6b9e67
- Vilhuber, L., Connolly, M., Koren, M., Llull, J., and Morrow, P. (2020), "A Template README for Social Science Replication Packages," Zenodo Version Number: v1.0.0.
- Vilhuber, L., Turitto, J., and Welch, K. (2020), "Report by the AEA Data Editor," *AEA Papers and Proceedings*, 110, 764–775.