Data Citations and Reproducibility in the Undergraduate Curriculum

Diego Mendez-Carbajo\textsuperscript{1} Alejandro Dellachiesa\textsuperscript{2}

\textsuperscript{1}Research Division, Federal Reserve Bank of St. Louis, St. Louis, Missouri, United States of America, \textsuperscript{2}Department of Economics, Gatton College of Business and Economics, University of Kentucky, Lexington, Kentucky, United States of America

Published on: Jul 27, 2023
DOI: \url{https://doi.org/10.1162/99608f92.c2835391}
License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
We believe reproducibility should be part of the undergraduate curriculum in economics because it is a valuable professional skill. It should be developed by consistently citing the data sources used in economic arguments. Good data citations help document the work of collecting the data used in research. We must instill the practice of citing the data both by leading by example and by enrolling the help of librarians. This is not a standalone skill associated with a single course or topic, but a fundamental skill for future academic and professional success. We argue it should be demonstrated and emphasized throughout the curriculum.

The scholarship of teaching and learning in economics documents multiple efforts to bring the quantitative dimension of our professional work closer to the undergraduate college curriculum. Economics educators describing data-focused assignments and projects (Halliday, 2019; Marshall & Underwood, 2019; Mendez-Carbajo, 2015, 2019; Wolfe, 2020; Wuthisatian & Thanetsunthorn, 2019) highlight the data-finding step of these projects. Even when the data sets are directly provided to the students (e.g., Easton, 2020), the instructors emphasize the broader data literacy dimensions of the assignments. However, there is neither professional consensus about how to build data literacy skills (Wuthisatian & Thanetsunthorn, 2019) nor much actual research on their mastery among economics students (Halliday, 2019).

In this essay, we discuss evidence of baseline proficiency levels among undergraduate college students related to identifying data series and their sources. We also put forward an accessible pedagogical strategy to develop basic reproducibility skills.

**Expected Proficiencies**

There is a natural overlap regarding the development of data literacy skills between economics and library science: both disciplines value it and contribute to its development.

The two seminal descriptions of expected proficiencies in data literacy among undergraduate students are provided by Hansen (2001) and Pothier and Condon (2019). The first of the seven broad competencies of economics majors named by Hansen directly addresses data provenance. It states: “Access existing knowledge: … Track down economic data and data sources. Find information about the generation, construction, and meaning of economic data” (p. 232).

The library science perspective provided by Pothier and Condon (2019) is articulated through seven expected data competencies of economics and business majors. The last one states: “Data ethics: The principles of data ethics are built on data ownership, intellectual property rights, appropriate attribution and citation, and confidentiality and privacy issues involving human subjects” (p. 139).

The utilitarian and ethical aspects of data reproducibility outlined above are emphasized by the profession in the American Economic Association’s (AEA; 2020) Data and Code Availability Policy, which clearly states “All source data used in the paper shall be cited, following the AEA Sample References.” However, the
scholarship documenting the collaboration in this area between instructional economics faculty and librarians is limited. Neither the calls by economics instructors (Li & Simonson, 2016; McGrath & Tiemann, 1985; Mendez-Carbajo, 2016) nor the experiences documented by librarians (Waggoner & Yates Habich, 2020; Wheatley, 2020; Wilhelm, 2021) regarding a need for data literacy appear to have broad impact.

**Evidence of Broad Data Literacy Skills**

Mendez-Carbajo (2020) documents baseline levels of data literacy competency in several areas key to the accurate and ethical use of data for communication and decision making among high school and college students.

In the [online economic education module](#) produced by the Federal Reserve Bank of St. Louis, “FRED Interactive: Information Literacy,” two separate groups of high school students (N= 450) and college students (N= 912) answer seven pretest questions. The questions are mapped to the data literacy competencies described by both Pothier and Condon (2019) and Hansen (2001).

The analysis finds effectively identical levels of average baseline data literacy competency between high school and college students. However, it also documents much higher levels of perceived self-efficacy among college students than among high school students. In other words, college students are no more knowledgeable or skilled than high school students but are significantly more confident in their work. This finding highlights a major challenge for instructors working to develop the expected proficiencies identified in the literature: the average college student is unduly comfortable in their limited understanding of the primary sources of economic data, and data literacy more broadly.

**Evidence of Narrow Reproducibility Skills**

During the fall semester of 2020, we distributed a short online assignment to all 854 students enrolled in two different upper-division economics courses offered by a large public university in the United States. The purpose of this assignment was to document the baseline competency of undergraduate college students to (i) identify the data, as well as their sources, referenced in an essay and to (ii) list the missing elements of a data citation. We consider these a small, yet foundational, set of reproducibility skills because good data citations help document the work of collecting the data used in research. We refer to them as ‘narrow’ skills because they are a subset of broader information literacy skills and complementary to other reproducibility skills such as code review.

On average, the students are slightly above 20 years of age, 49% identify themselves as female, 21% identify as non-White racial or ethnic minorities, and 92% report English is their native language. Academically, 87% of students are business, economics, or finance majors and hold a grade point average of 3.41. Also, 68% of students are currently enrolled in a statistics course required by their program and, on average, have previously completed more than one and a half economics courses.
The assignment had three sections:

- First, the students were directed to read a brief, 900-word essay on how to create data citations with FRED®. This essay provided background on the value of good data citations for practitioners of economics and could be used as reference material for the next two sections of the assignment.
- Second, the students were directed to read two short—under 600 words—economic essays. See them here and here. Each included a line graph of economic data. In the text, the authors referenced the data series and their sources while interpreting the quantitative information presented in the graph.
- Third, the students were asked to complete three tasks: identify the data series discussed in the essay; identify the sources of the data series discussed in the essay; and identify the missing elements of a data citation provided in the essay.

The assignment was completed in its entirety by 501 students. Table 1 reports our findings.

**Table 1. Data literacy skills.**

<table>
<thead>
<tr>
<th>Scores, Misconceptions, and Errors</th>
<th>Essay A</th>
<th>Essay B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score Correctly Identifying Series</td>
<td>0.57</td>
<td>0.47</td>
</tr>
<tr>
<td>Score Correctly Identifying Source</td>
<td>0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>Score Identifying Incomplete Citation</td>
<td>0.18</td>
<td>-0.04</td>
</tr>
<tr>
<td>Can't Identify Sources</td>
<td>0.05</td>
<td>0.12</td>
</tr>
<tr>
<td>Confuses Source with Distributor</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>Considers Citation to be Complete</td>
<td>0.25</td>
<td>0.40</td>
</tr>
</tbody>
</table>

The data literacy scores are calculated as:

\[
Score = \frac{(N\ Correct\ Answers - N\ Incorrect\ Answers)}{(N\ Correct\ Answers)}
\]

The scores can range between 1 (high skill, no incorrect answers) and −1 (low skill, no correct answers).

We document very weak student data literacy competencies associated with narrow reproducibility skills. Data literacy scores related to correctly identifying the sources of the data or recognizing an incomplete data citation are very low. Moreover, we document a frequent misconception of confusing the data source with the distributor.
These findings have practical implications for instructors, whether they are librarians or economic educators. Our work suggests there is a substantial instructional opportunity to help students develop the ability to recognize data series and their sources. In that regard, disambiguating the roles of data distributors and data sources can potentially yield large benefits to students, who would be able to acquire a more sophisticated understanding of how data are created and made available.

**Proposed Instructional Intervention**

We propose a broad instructional intervention for economics instructors reflecting the fact that correctly citing the data is a foundational literacy skill.

- Lead students by example and consistently name the sources of all data referenced or used in your teaching.
- Embed this practice in all your teaching, regardless of the subject or the type of course.
- Enroll the help of librarians by leveraging their ongoing instructional outreach on information literacy to include data citations.

Proficiency in identifying data sources is foundational to the development of reproducibility skills. The earlier and the more frequently students are exposed to best practices in data citations, the more effortlessly they will be able to adopt sophisticated professional replicability practices.

**Conclusion**

Reproducibility should be part of the undergraduate curriculum in economics. The buildup of data literacy skills is the first step to create a culture of reproducibility in the next generation of researchers. The scholarship of teaching and learning in both economics and library science already identifies citing data as an expected competency of economics students. The practice of teaching economics should embody that aspiration. To realize that aspiration, we put forward the following reflections and calls to action.

- Citing the sources of the data makes research work more thorough and possible to reproduce. It is a valuable professional skill that shows the background work that goes into doing economic research.
- This skill should be developed throughout the curriculum. This skill is not particular or exclusive to econometrics or statistics courses.
- The first step in teaching and instilling reproducibility in students is to develop citation skills for the data sources used in economic arguments. This includes data tables, plots, and in-text references.
- We must instill the practice by leading by example. Economics educators should enroll the help of librarians in developing this skill among students.
Disclosure Statement

The views expressed in this article are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

References


https://doi.org/10.1080/08963568.2020.1847551


©2023 Diego Mendez-Carbajo and Alejandro Dellachiesa. This article is licensed under a [Creative Commons Attribution (CC BY 4.0) International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.