Data Citations and Reproducibility in the Undergraduate Curriculum

Beyond the Numbers Conference
November 7, 2023
Authors

Diego Mendez-Carbajo, Ph.D.
Senior Economic Education Specialist
Federal Reserve Bank of St. Louis
diego.mendez-carbajo@stls.frb.org

Alejandro Dellachiesa
Lecturer in Economics
University of Kentucky
adellachiesa@uky.edu
Outline

• Framing the topic.
• Research design.
• Findings.
• Practical implications.
This information is my opinion and does not represent the official views of the Federal Open Market Committee, the Federal Reserve System or the Federal Reserve Bank of St. Louis.
Data Citations and Reproducibility in the Undergraduate Curriculum

by Diego Mendez-Carboajo and Alejandro Dellachiesa

https://doi.org/10.1162/99608f92.c2835391
Two Perspectives About Data literacy in Economics
LITERATURE REVIEW
Complementary perspectives

Economics

• Easton (2020); Wolfe (2020); Halliday (2019); Wuthisatian and Thanetsunthorn (2019); Marshall and Underwood (2019); Mendez-Carbajo (2015 & 2019).

Library science

• Wilhelm (2021); Wheatley (2020); Waggoner and Yates Habich (2020); Pothier and Condon (2019).
Economics
Hansen (2012)

Library Science
Pothier and Condon (2019)
American Economic Association’s Data and Code Availability Policy:
“All source data used in the paper shall be cited” (2019)
Economic Data Literacy Skills Among High School and Undergraduate Students

EMPIRICAL EVIDENCE
Baseline Competency and Student Self-efficacy in Data Literacy: Evidence from an Online Module

Diego Mendez-Carbajo

https://doi.org/10.1080/08963568.2020.1847551
Evidence from an Online Module

FRED Interactive Module *Information Literacy*.

- Produced by the Research Information Services team from the Federal Reserve Bank of St. Louis.
- Seven pre-test questions, mapped to the data literacy competencies described by Pothier and Condon (2019)
  - High school students (*N* = 450)
  - College students (*N* = 912)
Figure 1: Percent of correct answers and confidence by pre-test question and student type
New Research Questions

• **How skilled** are undergraduate students at identifying the data used in economic arguments?

• **How knowledgeable** about economic data sources are undergraduate students?

• **Can undergraduate students identify** a complete data citation?

• **Do baseline data literacy skills impact the perception of economic research being reproducible?**
Assignment and Survey

INSTRUMENTS AND METRICS
Data Literacy Assignment

• Three sections.
  – Instruction: Descriptive essay on citations.
  – Practice: Two economic letters and summative assessments.
  – Reflection: Compare and report perceived reproducibility and replicability.
Assignment: Instruction
**Essay A**

**Gasoline Affordability**

William T. Garing

In February 1999, the average production worker in the United States earned $13.28 per hour, enough to buy more than 200 gallons of gasoline, which, according to a Department of Energy nationwide survey, was selling at $4.92 per gallon. By May 2004, the average hourly wage had risen about 18 percent to $15.63 per hour, but the price of gasoline had more than 100 percent to $3.98 per gallon. Thus, an hour of work in May would purchase less than 8 gallons of gasoline. Gasoline’s increased cost has led some to speculate that Americans will lose their appetite for gas guzzling SUVs.

February 1999, however, was the low point in the history of gasoline prices relative to hourly earnings. The average worker at that time could purchase more gasoline with an hour’s wage than in any month going back to 1967. Further, May 2004 is far from the high point in gasoline costs. In March 1981, the hourly wage was $7.28 and the price of gasoline was about $1.30 per gallon. The Department of Energy survey data on retail gasoline prices does not begin until 1999, but we do have the Bureau of Labor Statistics Consumer Price Index (CPI) on the average price of gasoline. This index was 263.1 in January 1967—when the average worker was paid $2.78 per hour—and rose to 113 in March 1981. The same index was 85.1 in February 1999 and 162.2 in May 2004. So the actual price paid at the pump was about 25 percent lower in 1999 than it was in 1981, and the wage rate was almost twice as high.

The chart presents an index of the cost of gasoline relative to the average hourly earnings of production workers in the United States. It is the ratio of the CPI index for the price of gasoline divided by the average hourly wage rate. During the past 50 years, the cost of gasoline relative to the wage rate has fallen, with wide fluctuation around the trend. The chart includes a trend line equal to the 1967 ratio—time when Americans did not worry much about fuel efficiency.

**Essay B**

**Renewable Sources of Electricity: Where Excess Capacity Is Built-In**

Teresa Mendez-Carballo, Senior Economic Education Specialist

Although electricity can be generated in multiple ways, it is costly and impractical to store electricity in large amounts. When there is high demand for electricity, for example, during a hot day when air conditioners run for hours, electricity must be produced rather than taken from storage facilities. After words, consumption and production of electricity generally move in lock-step. If there is not enough electricity, plants stop working. But minimizing production capacity at the ready for when there is more demand is an expensive investment for utility companies. The challenge for utility companies is to provide energy at low costs for uncertain and variable demand.

In the figure, data from the Survey of Industrial Activity by the Board of Governors of the Federal Reserve System show annual electricity production (blue line) and capacity utilization (red line) between 1982 and 2019. Electricity production is measured by an index, which is equal to 100 in 1982 and 2012, and capacity utilization is measured as the percent of total electricity production capacity that is actually put to use, on average. Between 1982 and 2006, production grew by 73 percent and average capacity utilization increased from 80 percent to 97 percent. In other words, as the demand for and the supply of electricity increased, the unused production capacity of electricity decreased. Utilities bought more power than their maximum production capacity. Between 2001 and 2019, this trend reversed. Electricity production decreased an additional 14 percent while capacity utilization decreased from 92 percent to 75 percent. Utilities operated with greater spare capacity from 2001 to 2019.

As renewable sources of electricity have expanded, production capacity utilization has gradually decreased.

The development of renewable electricity sources—for example, solar and wind—might help explain this increase in spare capacity from 2001 to 2019. The rapid expansion of solar parks and wind turbine farms has made those methods of generating electricity the largest renewable source of electricity in the United States. In 2019, the combined output surpassed hydroelectric production in the United States.
Assignment: Practice

Identify:

- Data used in the essay.
- Sources of data used in the essay.
- Missing elements of an incomplete data citation.
Assignment: Reflection

Compare the data citations in both essays. Based on the data citations, which data analysis is easier for you to reproduce and replicate?

- Essay A: “Gasoline Affordability”
- Essay B: “Renewable Sources of Electricity”
- Both essays are equally difficult to replicate.
- Both essays are equally easy to replicate.
Survey: Student Characteristics

• **Demographics**: age; gender; race or ethnicity; and native language.

• **Academic**: grade point average; declared major; concurrent enrollment in a statistics course required by their program; and number of economics course already completed.
Metrics

• Data Literacy Scores:

\[
Score = \frac{\#\text{ Correct Answers} - \#\text{Incorrect Answers}}{\#\text{ Correct Answers}}
\]

• Misconceptions and errors:
  – Confusing data distributor and data source.
  – Considering a data citation is complete.
Descriptive Statistics

DATA
## Population and sample sizes

<table>
<thead>
<tr>
<th>Population and Sample</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Population</td>
<td>854</td>
</tr>
<tr>
<td>Participants</td>
<td>661</td>
</tr>
<tr>
<td>Started Assignment</td>
<td>519</td>
</tr>
<tr>
<td>Completed Assignment</td>
<td>501</td>
</tr>
</tbody>
</table>
# Student profile

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>20.43</td>
</tr>
<tr>
<td>Female</td>
<td>0.49</td>
</tr>
<tr>
<td>Minority</td>
<td>0.21</td>
</tr>
<tr>
<td>Native Language is English</td>
<td>0.92</td>
</tr>
<tr>
<td>Bus/Econ/Finance Major</td>
<td>0.87</td>
</tr>
<tr>
<td>Grade Point Average</td>
<td>3.41</td>
</tr>
<tr>
<td>Statistics Course</td>
<td>0.68</td>
</tr>
<tr>
<td>Previous Economics Courses</td>
<td>1.63</td>
</tr>
</tbody>
</table>
Self-Selection and Demonstrated Skills

FINDINGS
### Dependent variable

<table>
<thead>
<tr>
<th>Probability of starting assignment (N= 519)</th>
<th>Probability of completing assignment (N= 501)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constant</strong></td>
<td></td>
</tr>
<tr>
<td>-2.4986 ***</td>
<td>0.1404</td>
</tr>
<tr>
<td>(-2.38)</td>
<td>(0.18)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>0.0718 *</td>
<td></td>
</tr>
<tr>
<td>(1.73)</td>
<td></td>
</tr>
<tr>
<td><strong>College GPA</strong></td>
<td></td>
</tr>
<tr>
<td>0.4244 ***</td>
<td>0.5004 **</td>
</tr>
<tr>
<td>(3.27)</td>
<td>(2.12)</td>
</tr>
<tr>
<td><strong>Statistics Course</strong></td>
<td></td>
</tr>
<tr>
<td>0.6180 ***</td>
<td></td>
</tr>
<tr>
<td>(5.26)</td>
<td></td>
</tr>
<tr>
<td><strong>McFadden R-squared</strong></td>
<td></td>
</tr>
<tr>
<td>0.0617</td>
<td>0.0283</td>
</tr>
</tbody>
</table>

Note: Asterisks denote the significance level of the z-statistic (*** 0.01, ** 0.05, * 0.1)
<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:

\[
\text{Score} = \frac{(N_{Correct \, \text{Answers}} - N_{Incorrect \, \text{Answers}})}{(N_{Correct \, \text{Answers}})}
\]

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

INSTRUCTION AND CURRICULUM
Recommendations for Instructors

• **Enroll** the help of librarians.
• **Consistently** name the sources of data.
• **Embed** the practice in all your teaching.
• **Lead** by example.
Citing the data is a foundational skill: Practice it across the curriculum.
Questions?