FEDERAL RESERVE BANK of ST. LOUIS
CENTRAL TO AMERICA'S ECONOMY®

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.

HARVARD DATA SCIENCE REVIEW

https://doi.org/10.1162/99608f92.c2835391

Two Perspectives About Data literacy in Economics

LITERATURE REVIEW

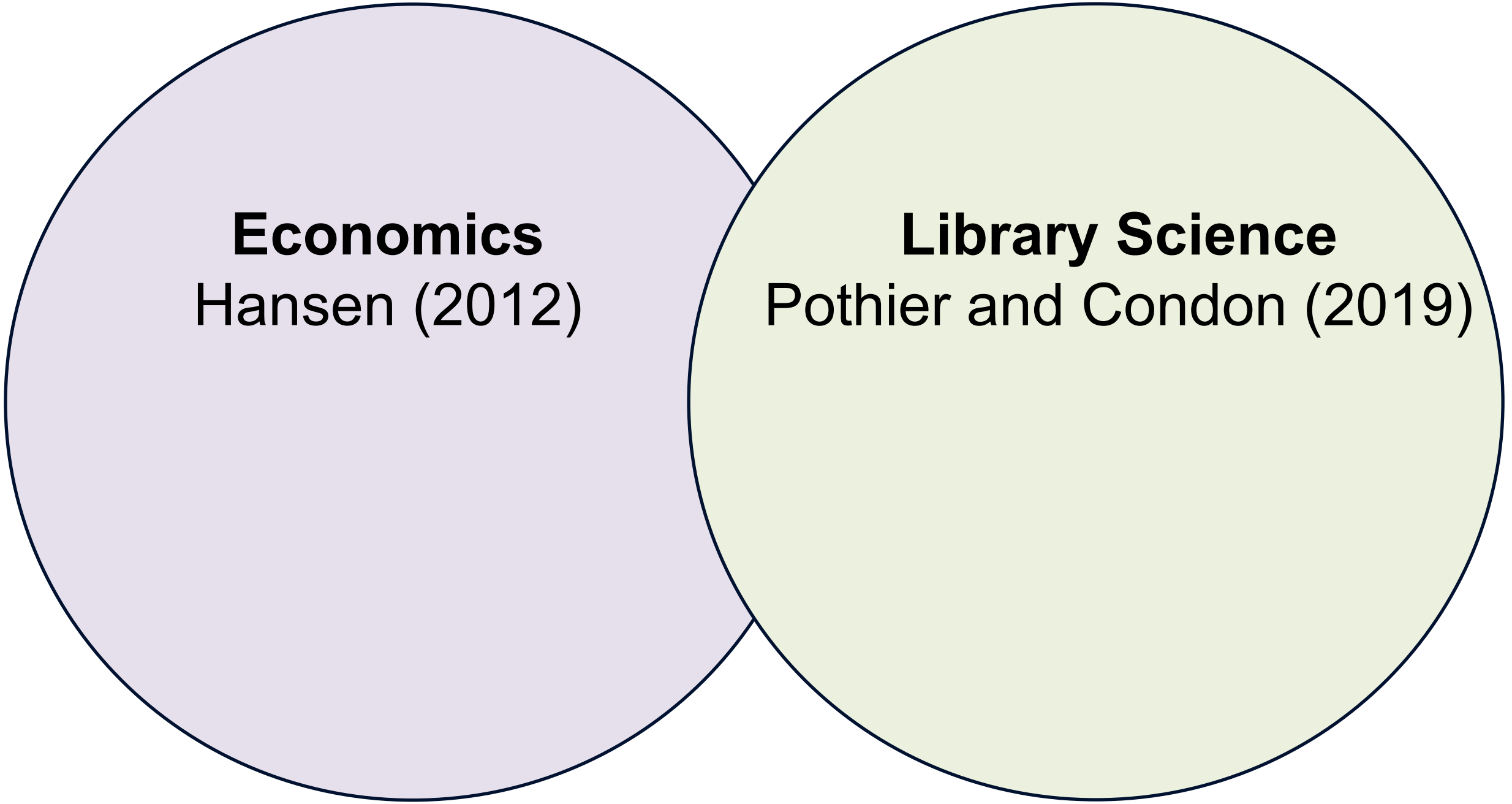
Complementary perspectives

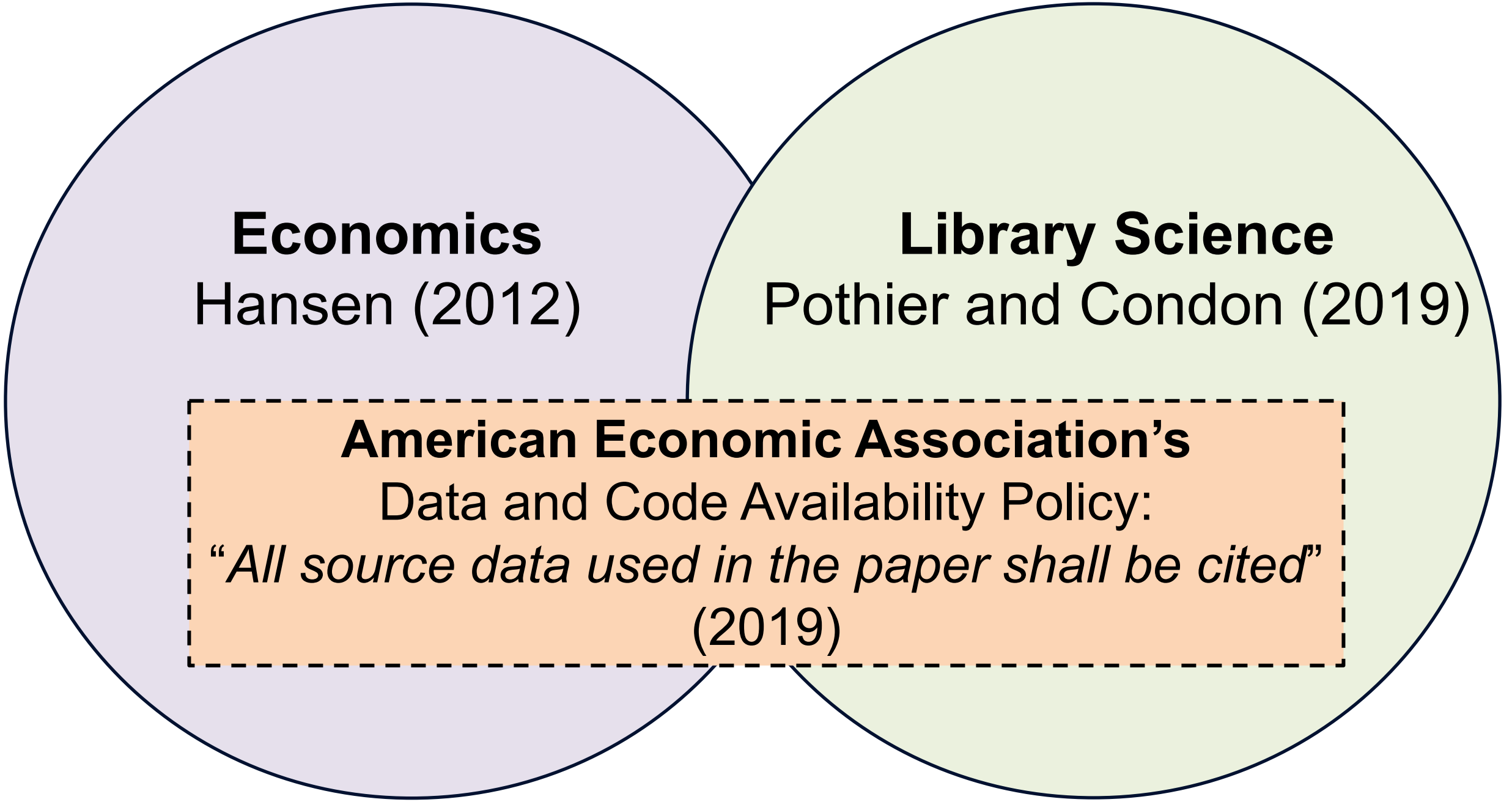
Economics

- 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)


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

Routledge
Taylor & Francis Group

Journal of Business & Finance Librarianship

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/wbfl20

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

Diego Mendez-Carbajo

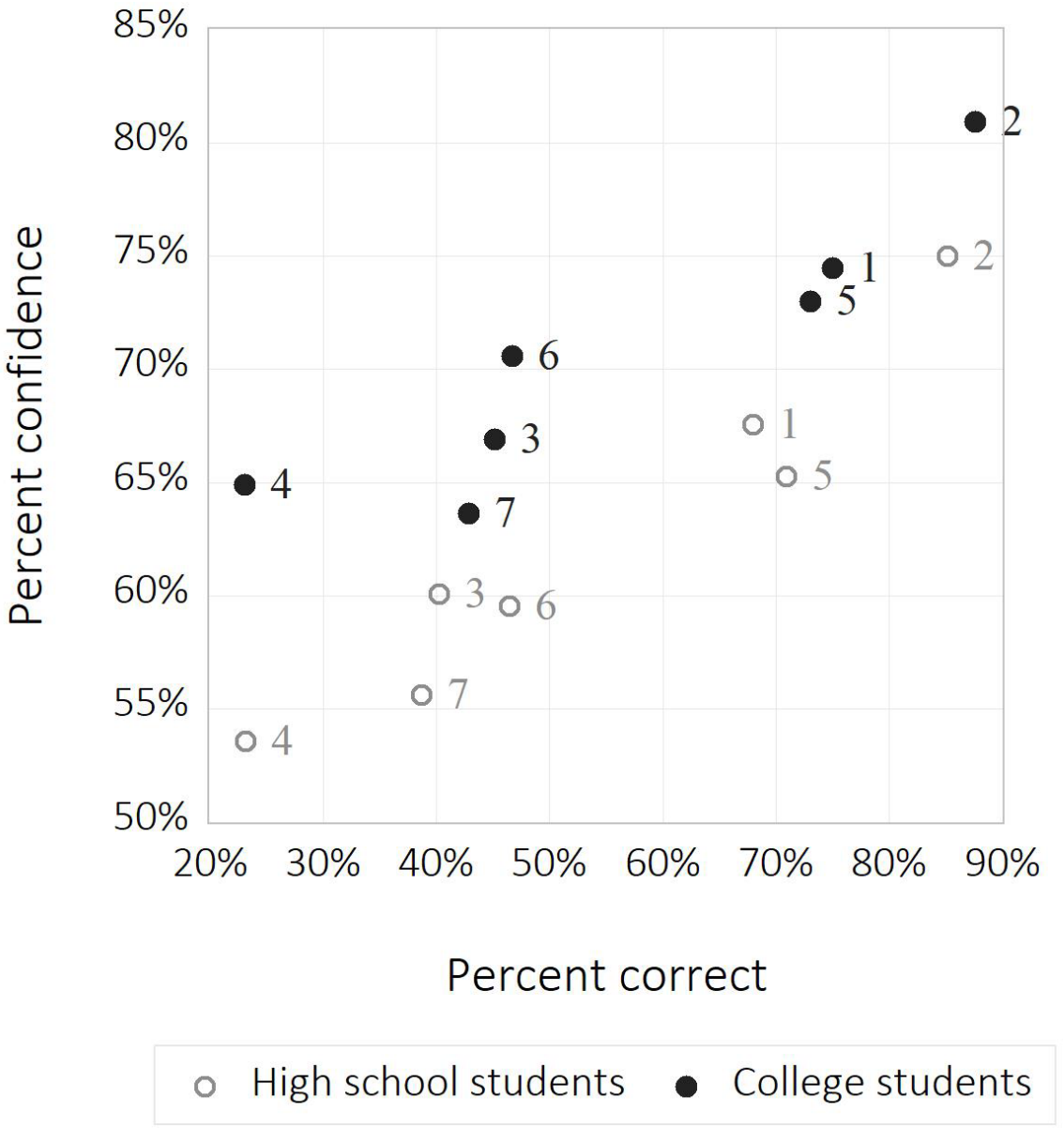
https://doi.org/10.1080/08963568.2020.1847551

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?**

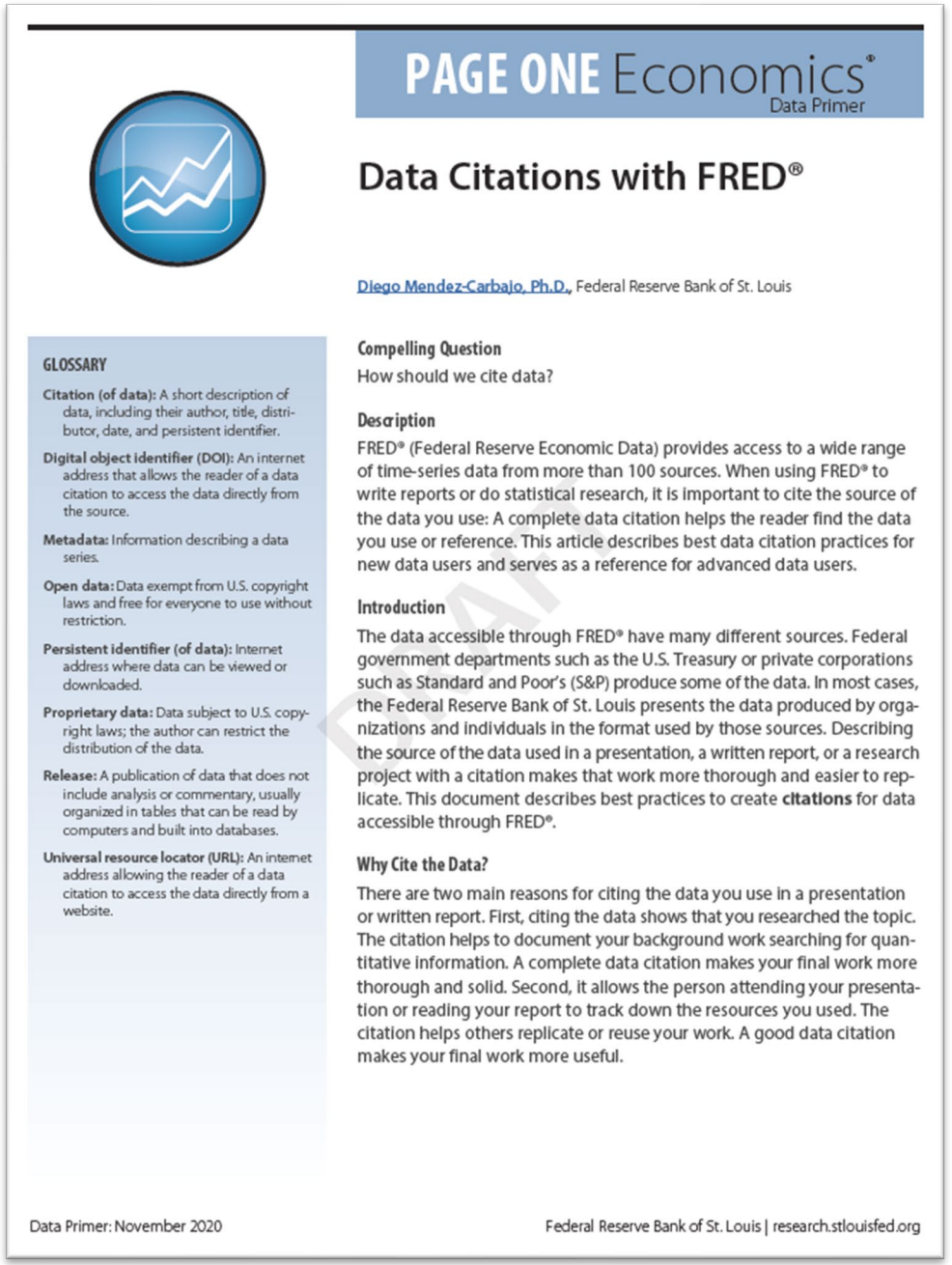
Assignment and Survey

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



PAGE ONE Economics®
Data Primer

Data Citations with FRED®

Diego Mendez-Carbajo, Ph.D., Federal Reserve Bank of St. Louis

GLOSSARY

Citation (of data): A short description of data, including their author, title, distributor, date, and persistent identifier.

Digital object identifier (DOI): An internet address that allows the reader of a data citation to access the data directly from the source.

Metadata: Information describing a data series.

Open data: Data exempt from U.S. copyright laws and free for everyone to use without restriction.

Persistent identifier (of data): Internet address where data can be viewed or downloaded.

Proprietary data: Data subject to U.S. copyright laws; the author can restrict the distribution of the data.

Release: A publication of data that does not include analysis or commentary, usually organized in tables that can be read by computers and built into databases.

Universal resource locator (URL): An internet address allowing the reader of a data citation to access the data directly from a website.

Compelling Question

How should we cite data?

Description

FRED® (Federal Reserve Economic Data) provides access to a wide range of time-series data from more than 100 sources. When using FRED® to write reports or do statistical research, it is important to cite the source of the data you use: A complete data citation helps the reader find the data you use or reference. This article describes best data citation practices for new data users and serves as a reference for advanced data users.

Introduction

The data accessible through FRED® have many different sources. Federal government departments such as the U.S. Treasury or private corporations such as Standard and Poor's (S&P) produce some of the data. In most cases, the Federal Reserve Bank of St. Louis presents the data produced by organizations and individuals in the format used by those sources. Describing the source of the data used in a presentation, a written report, or a research project with a citation makes that work more thorough and easier to replicate. This document describes best practices to create **citations** for data accessible through FRED®.

Why Cite the Data?

There are two main reasons for citing the data you use in a presentation or written report. First, citing the data shows that you researched the topic. The citation helps to document your background work searching for quantitative information. A complete data citation makes your final work more thorough and solid. Second, it allows the person attending your presentation or reading your report to track down the resources you used. The citation helps others replicate or reuse your work. A good data citation makes your final work more useful.

Essay A

Economic SYNOPSES
short essays and reports on the economic issues of the day
2004 ■ Number 27

Gasoline Affordability

William T. Gavin

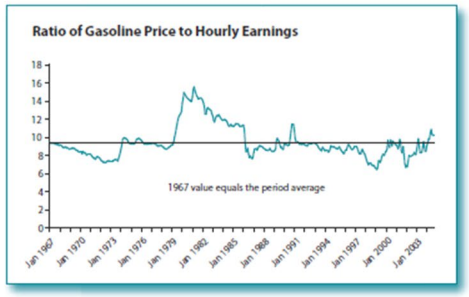
In February 1999, the average production worker in the United States earned $13.28 per hour, enough to buy more than 14 gallons of gasoline, which, according to a Department of Energy nationwide survey, was selling at $0.92 per gallon. By May 2004, the average hourly wage had risen about 18 percent to $15.63 per hour, but the price of gasoline had risen more than 100 percent to $1.98 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. Furthermore, 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 1990, but we do have the Bureau of Labor Statistics Consumer Price Index (CPI) on the average price of gasoline. This index was 26.3 in January 1967—when the average worker was paid $2.79 per hour—and rose to 113 in March 1981. The same index was 85.1 in February 1999 and 164.2 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 38 years, the cost of gasoline relative to the wage rate has been flat, with wide fluctuation around the trend. The chart includes a trend line equal to the 1967 ratio—a time when Americans did not worry much about fuel efficiency.

Between 1967 and 1973, the price of gasoline was relatively stable while wages rose, making gasoline more affordable for the average worker. The 1973 oil price hike led to a rapid rise in gasoline prices in 1974. Afterward, gasoline costs remained relatively flat with a slight downward trend until the next oil price shock hit in 1979. Our index shows that the cost of gasoline relative to a worker's hourly wage reached a peak in March 1981, declined sharply in 1986, and remained relatively stable for the next decade. There was a brief spike in 1990 when Iraq occupied Kuwait, but the price quickly settled back to the 1967 norm.

In 1997 and 1998, falling oil prices led to a decline in gasoline prices and to a peak in the affordability of gasoline in early 1999. Since then, gasoline prices have become more volatile, but they have not strayed far from the affordability level that we saw in 1967. Unless this modestly higher price persists and continues to rise in tandem with or faster than wages, we would not expect it to dent consumer demand for SUVs. ■



Essay B

2020 ■ Number 42
https://doi.org/10.20955/es.2020.42
ECONOMIC Synopses

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

Diego Mendez-Carbajo, Senior Economic Education Specialist

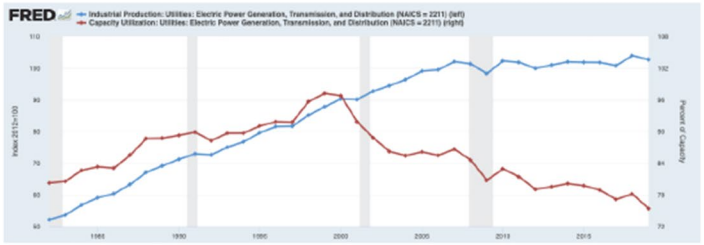
Although electricity can be generated in multiple ways, it is costly and impractical to store electricity in large amounts. When there is high demand for electricity, for example, during a hot day when air conditioners run for hours, electricity must be produced rather than taken from storage facilities. In other words, consumption and production of electricity generally move in lockstep.[1] If there is not enough electricity, appliances stop working. But maintaining production capacity at the ready for when there is more demand is an expensive investment for utility companies.[2] The challenge for utility companies is to provide energy at low costs for uncertain and variable demand.[3]

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 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 2000, 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 operated closer to their maximum production capacity. Between 2001 and 2019, this trend reversed. Electricity production increased 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 eolic—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, their combined output surpassed hydroelectric production by

Industrial Production and Capacity Utilization: Electric Power Generation, Transmission, and Distribution (1982-2019)



NOTE: Gray bars indicate recessions as determined by the National Bureau of Economic Research.
SOURCE: Board of Governors of the Federal Reserve System and FRED®, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/graph/?g=uXJp

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:

$$Score = \frac{\#\ Correct\ Answers\ -\ \#Incorrect\ Answers}{\#\ Correct\ Answers}$$

- Misconceptions and errors:
  - Confusing data distributor and data source.
  - Considering a data citation is complete.

Descriptive Statistics

DATA

Population and sample sizes

Population and Sample	*N*
Target Population	854
Participants	661
Started Assignment	519
Completed Assignment	501

Student profile

	Mean
Age	20.43
Female	0.49
Minority	0.21
Native Language is English	0.92
Bus/Econ/Finance Major	0.87
Grade Point Average	3.41
Statistics Course	0.68
Previous Economics Courses	1.63

Self-Selection and Demonstrated Skills

FINDINGS

	Dependent variable	
	Probability of starting assignment (*N*= 519)	**Probability of completing assignment** (*N*= 501)
Constant	-2.4986 *** (-2.38)	0.1404 (0.18)
Age	0.0718 * (1.73)	
College GPA	0.4244 *** (3.27)	0.5004 ** (2.12)
Statistics Course	0.6180 *** (5.26)	
McFadden R-squared	0.0617	0.0283

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

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

The data literacy scores are calculated as:

$$Score = (N_{Correct\ Answers} - N_{Incorrect\ Answers}) / (N_{Correct\ Answers})$$

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

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?