Replication Code Availability Over Time and Across Fields: Evidence from the SOEP

Lukas Fink & Jan Marcus

Freie Universität Berlin

Leibniz Open Science Day 2024: Meta-Perspectives in Social Sciences

November 25, 2024

This project I

- (Computational) **reproducibility**: "the result obtained by running the original code on the original data equals the original result" (Pérignon et al. 2023: 3)
- Two crucial ingredients: code and data
- We focus on code sharing
  - Everyone can share code
  - Not everyone is allowed to share data (for legal and privacy reasons)

This project I

- (Computational) **reproducibility**: "the result obtained by running the original code on the original data equals the original result" (Pérignon et al. 2023: 3)
- Two crucial ingredients: code and data
- We focus on code sharing
  - Everyone can share code
  - Not everyone is allowed to share data (for legal and privacy reasons)

- Extent of code sharing among social scientists is unclear (despite the relevance of code sharing for reproducibility)

  $\Rightarrow$ First study on extent of code sharing and its development over time
- We focus on a setting in which the data are available for researchers worldwide and free of charge (SOEP).

  $\Rightarrow$ Provision of code is the crucial factor for reproducibility

Research questions

- What factors correlate with code provision?
- How has code provision evolved over time and what are potential explanations for this development?
- Is code provision related to quality metrics (e.g., impact factors, citations)?

Research questions

- What factors correlate with code provision?
- How has code provision evolved over time and what are potential explanations for this development?
- Is code provision related to quality metrics (e.g., impact factors, citations)?

→ Note: These are descriptive research questions.

Why is code provision important?

It strengthens trust in scientific results & integrity of the scientific process

- Helps counteracting the "replication/reproducibility crisis"
- Facilitates discovery and correction of coding errors
  - $\Rightarrow$ Self-correction feature of scientific system
- Serves as a deterrent against fraud and questionable research practices
  - e.g., p-hacking

Why is code provision important? II

Because it allows to build on prior research more easily

- Enables further investigation of the subject matter, e.g.
  - further outcomes
  - effect heterogeneity
  - additional robustness exercises
  - Think of the case, where different researchers come to different conclusions for the same research question
  - More sophisticated meta-analyses
- Enhances knowledge dissemination by providing insights into applied methods and coding practices

Outline

2. Data

1. SOEP & SOEPlit

The German Socio-Economic Panel (SOEP)

- One of the largest and longest-running household panel surveys worldwide.
- First interviews conducted 40 years ago (in 1984).
- Approximately 30,000 individuals in about 15,000 households are annually surveyed.
- Covers a wide range of topics:
  - Employment, earnings, health, education, demography, income, housing.
  - Life satisfaction, attitudes, values, and personality.

Advantages of the SOEP for our research question

1. Researchers worldwide have free and open access to SOEP data.
   - Admin data: costly + restricted access environments.

   $\Rightarrow$ Replications with the SOEP should be more easily feasible.
2. Clear separation of the data collector from the data analyst.
   - No "strategic" decisions during data collection to influence direction and significance of results.
3. Widely used by the scientific community.
   - More than 3,500 users worldwide.
   - Over 2,500 peer-reviewed, SOEP-based journal articles have been published.
4. Long time horizon of publications
5. Used by researchers from various disciplines
6. List of all SOEP publications allows to clearly define target population

Study population: SOEPlit

- A database maintained by the SOEP group at DIW
- Aim: Covering all publications based on SOEP data.
- Includes bibliographic information about each publication:
  - Title, type of publication, publication year, language.
  - Journal (if applicable), digital object identifier (DOI, if available), and author names.
- Analyses focus on peer-reviewed articles published in journals listed in the relevant citation indices:
  - Social Science Citation Index (SSCI), Science Citation Index (SCI), Arts and Humanities Citation Index (AHCI), Emerging Sources Citation Index (ESCI).
- (We exclude working papers to avoid double-counting and due to potential author reluctance to share code for working papers.)

$\Rightarrow$ Population of **N=2,518** unique publications from 1985 to 2021.

2. Data

2. Variables

Main variable: Code availability

We checked code availability (yes/no) for each publication in several ways:

1. Journal webpage.
2. Authors' webpages.
3. Specific online repositories.
   - GESIS data archive, Harvard Dataverse, Open Science Framework, openICPSR, Zenodo.
4. Articles themselves.
   - Student assistants checked acknowledgments and searched for key terms (code, replication, syntax, Stata).

Main variable: Code availability

We checked code availability (yes/no) for each publication in several ways:

1. Journal webpage.
2. Authors' webpages.
3. Specific online repositories.
   - GESIS data archive, Harvard Dataverse, Open Science Framework, openICPSR, Zenodo.
4. Articles themselves.
   - Student assistants checked acknowledgments and searched for key terms (code, replication, syntax, Stata).

$\rightarrow$ We verified that the code is accessible by downloading and opening it

Further information

We mapped additional information into our dataset

- **Article-level information**
  - Citations in Google Scholar

  → Source: Web scraping using the Scholarly package in Python (Cholewiak et al. 2021)

- **Journal-level information**
  - Primary discipline

  → Source: based on Science-Metrix Most frequent econ Most frequent others
  - Journal impact factor (JIF) metrics (since 1997)

  → Source: Master Journal List from Clarivate
  - Policies on replication code availability (see below)

2. Data

3. Descriptive statistics

Descriptive statistics

	Mean	SD	Min	Max
Economics	0.45	0.50	0.00	1.00
Sociology	0.16	0.36	0.00	1.00
Psychology	0.11	0.31	0.00	1.00
Other Social Sciences	0.15	0.36	0.00	1.00
Health and Other Sciences	0.14	0.34	0.00	1.00
Publication Year	2011.91	7.38	1985.00	2021.00
Journal Impact Factor (JIF)	1.86	2.20	0.04	40.14
5 Year JIF	2.60	2.58	0.09	43.77
Citation Count from Google Scholar	106.19	242.25	0.00	4359.00
Language of Article is English	0.87	0.33	0.00	1.00
Single Author	0.28	0.45	0.00	1.00
Two Authors	0.38	0.49	0.00	1.00
Three Authors	0.21	0.41	0.00	1.00
Four or More Authors	0.13	0.33	0.00	1.00

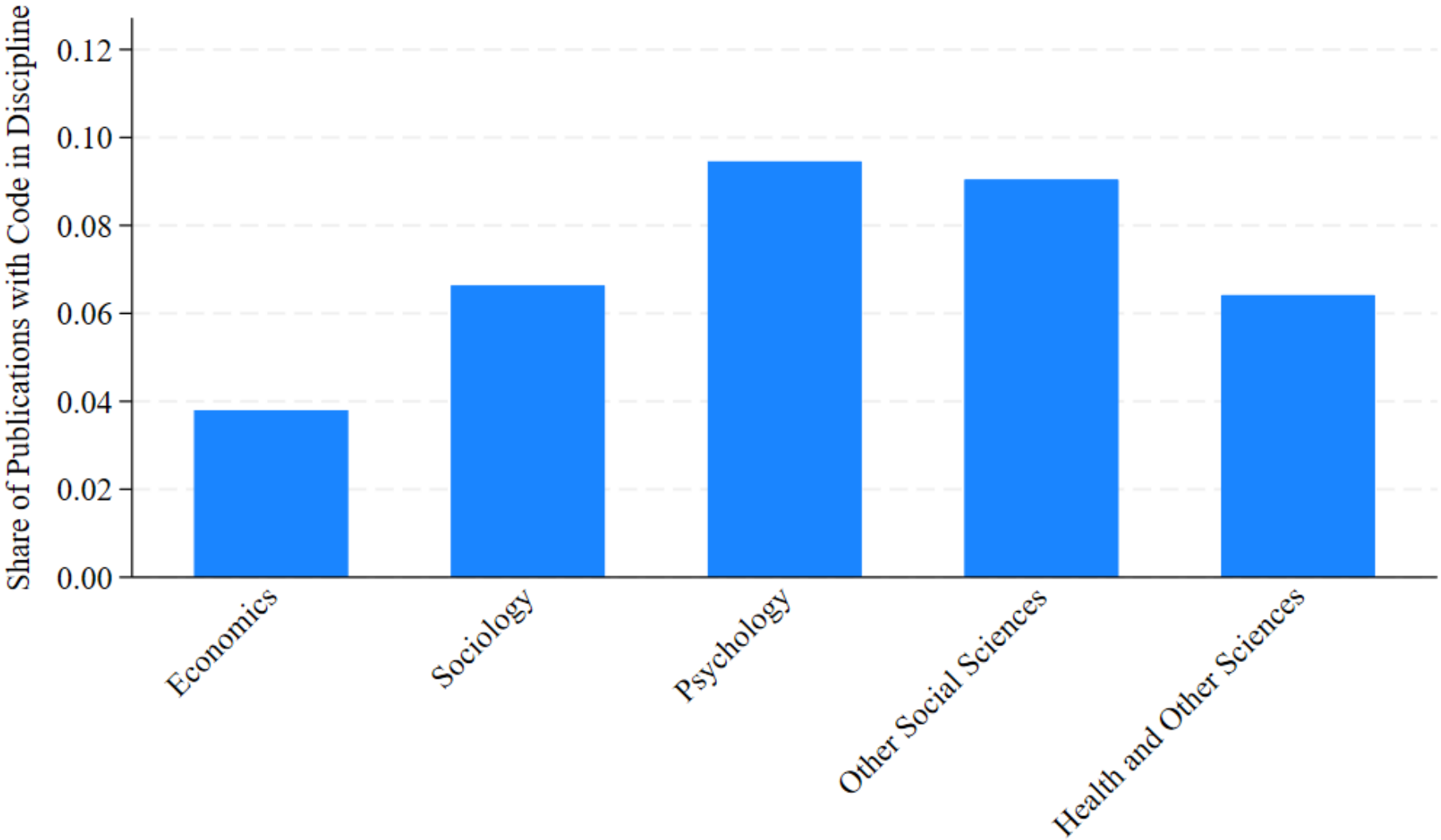
3. Results on replication code availability

1. The state of replication code availability

Replication code availability

- Replication code found for 151 of the 2,518 SOEP-based publications (6% across all disciplines).
  - 3.8% in economics journals (43 of 1,132)
  - 7.8% in journals of other disciplines (108 of 1,386).

Code availability by discipline



Further correlates & more detailed examination

Correlates of code availability

- \- Single-author article details
- \+ English-language article details

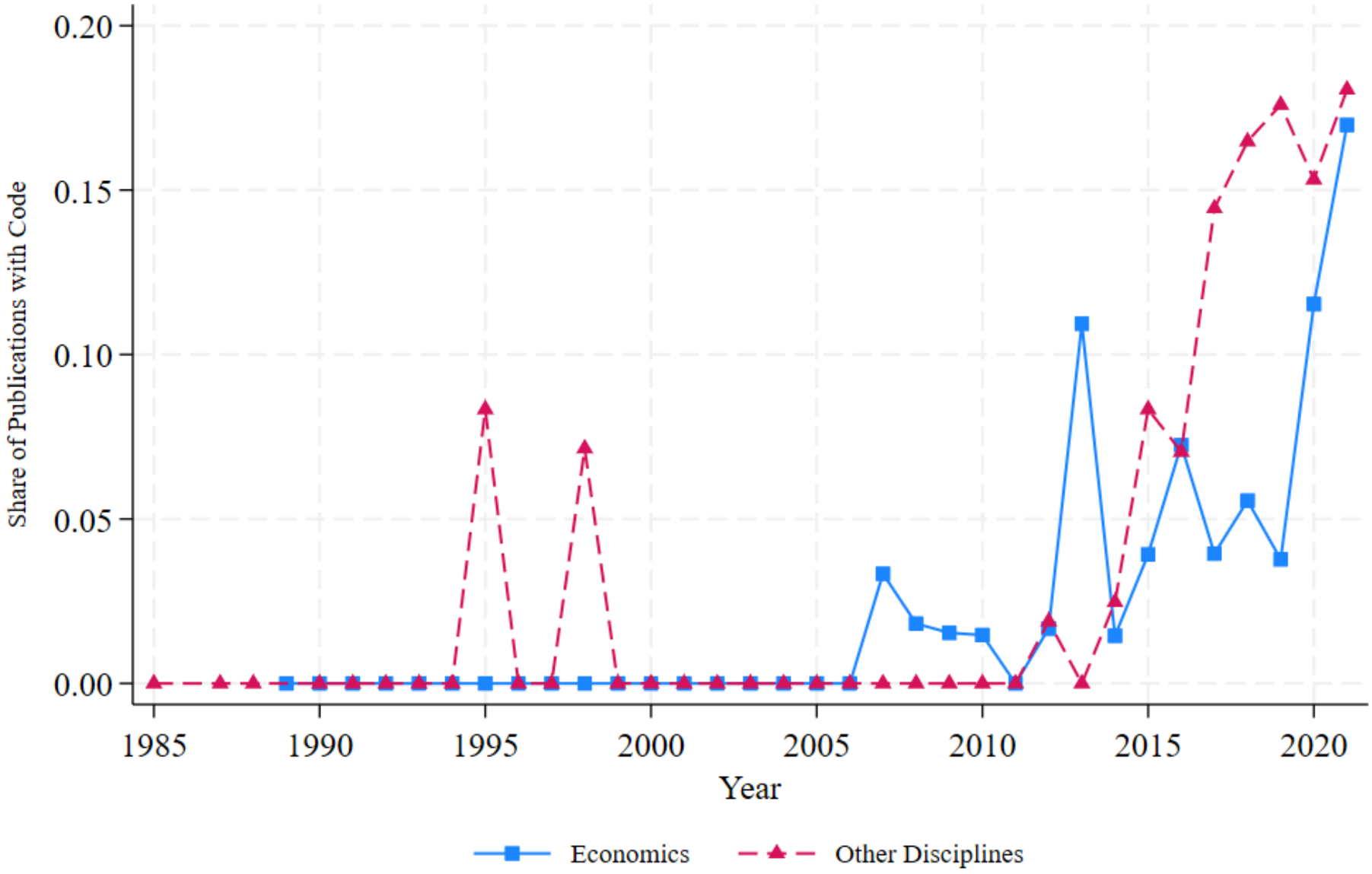
Examining studies with code

- **Software**: Code is mostly in Stata (85% of articles with code), followed by R (20 %); SPSS, Matlab, SAS all have less then 5% Graph
- **Mode of disclosure**: Mostly through journal details
  - However, for 32% in economics and 18% in other disciplines code availability was only disclosed via the authors' websites
- **Storage location**: Mostly through direct download (economics), other disciplines often through Open Science Framework (OSF) details

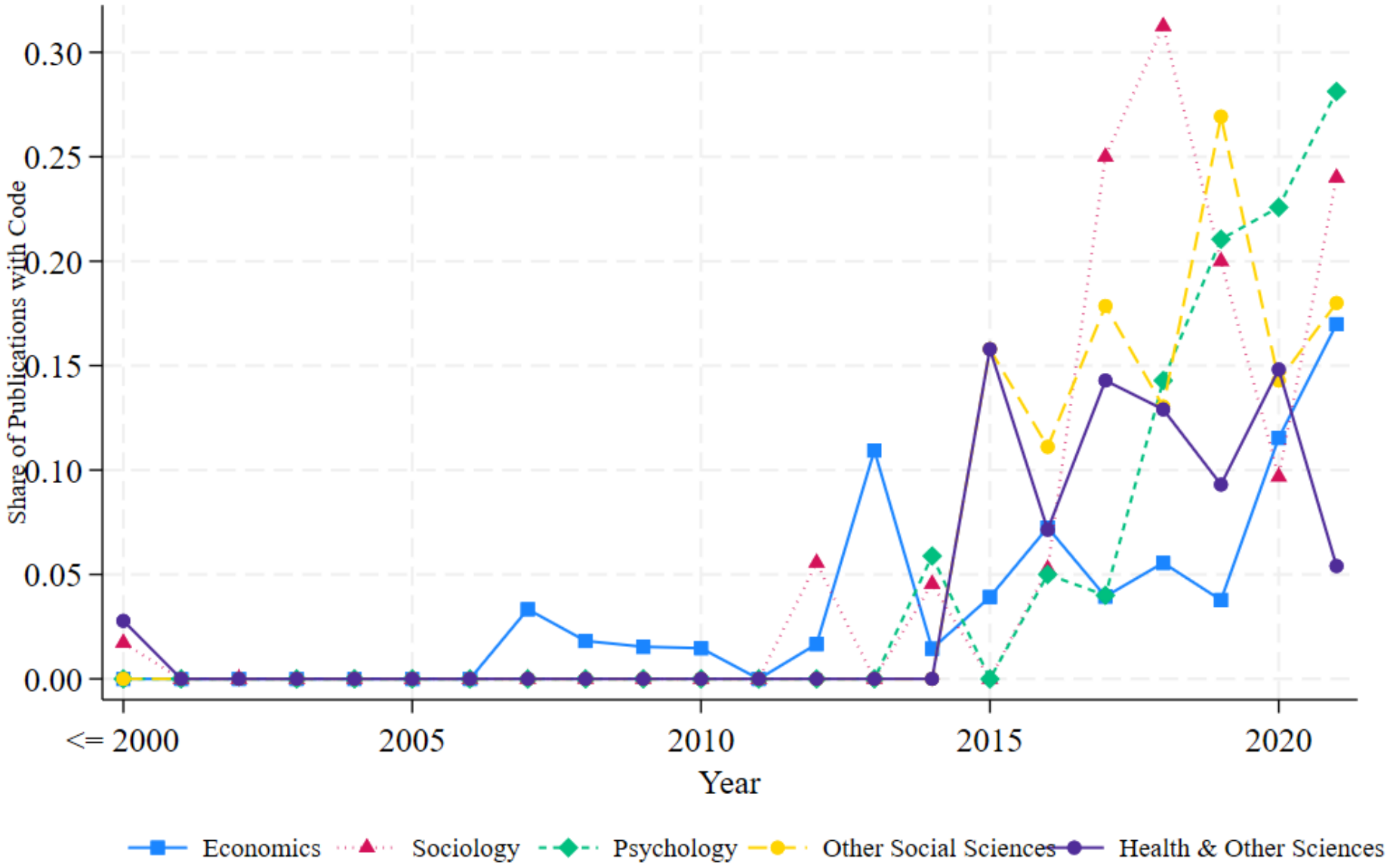
3. Results on replication code availability

2. Developments in code availability over time

Code availability over time



Code availability over time, across disciplines



Potential explanations for increased code availability

1. Technological advances
2. Journal policies
3. Awareness of researchers

1. Technological advances

- Widespread use of individual websites by researchers
- Establishment of journal websites
- Creation of specialized online repositories (e.g., OSF in 2012, Harvard Dataverse Repository in 2006)

$\Rightarrow$ Financial and logistic costs of code sharing strongly decreased

- Storage and exchange of floppy disks no longer required (as well as printing datasets and codes in the appendix)

Empirical support

- Increased use of repositories to store code
- Many studies for which code is available on the authors' websites.

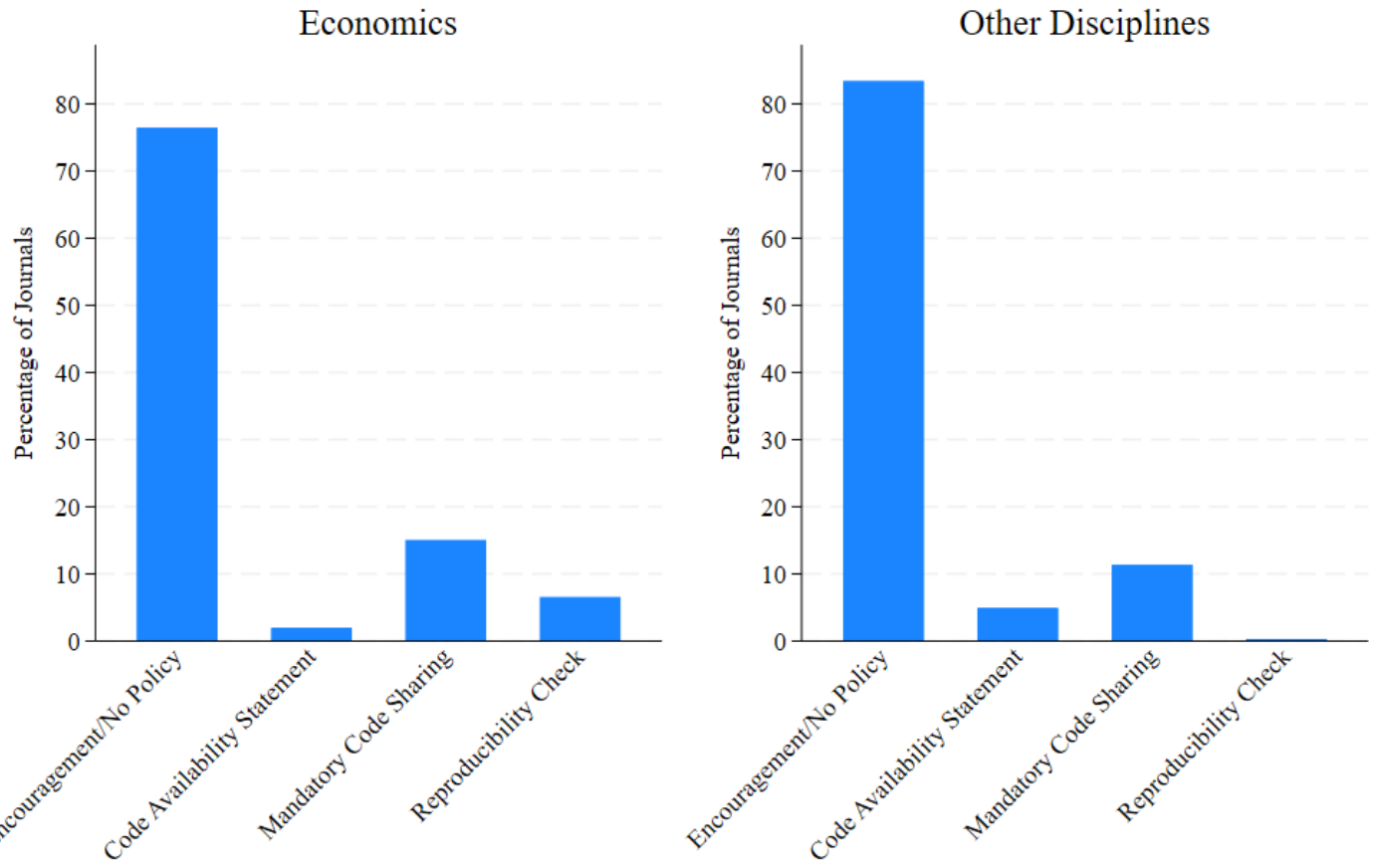
2. Journal policies on code sharing

Transparency and Openness Promotion (TOP) categorization for replication code policies

- **Encouragement (Level 0):** Journal encourages code sharing, or says nothing.
- **Code availability statement (Level 1):** Article states whether code is available, and, if so, where to access it.
- **Mandatory code sharing (Level 2):** Code must be posted to a trusted repository. Exceptions must be identified at article submission.
- **Reproducibility check (Level 3):** Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.

→ See Nosek et al. (2015) and the TOP Factor Database

Journal policies: Current state (2023) in our sample



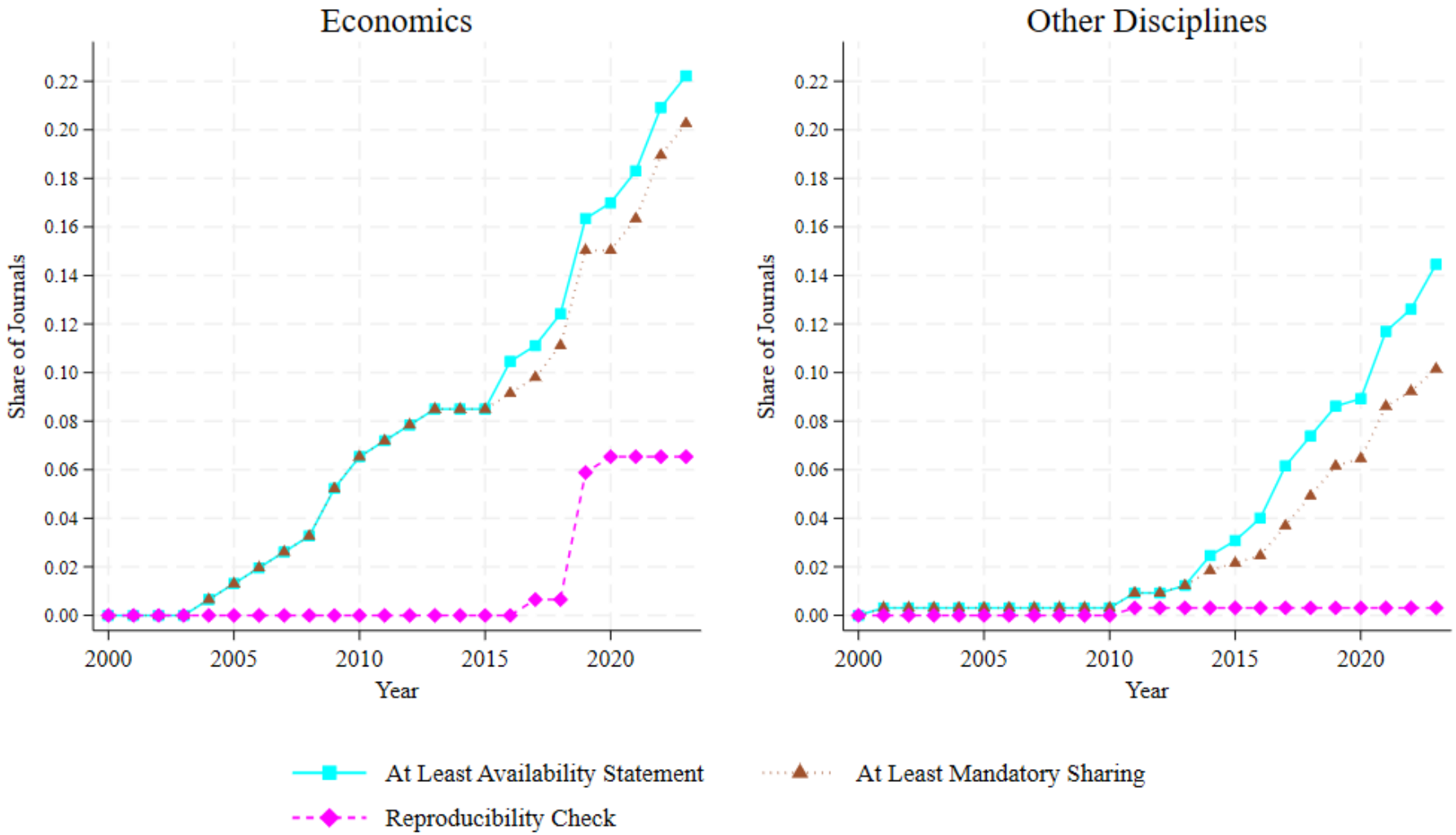
- **TOP Factor Database:** Includes the *current* state of journal policies for 164 of the 478 journals (34%) in our sample.
- We searched the websites of the other 315 journals for information on *current* code sharing policies

Historical development of journal policies

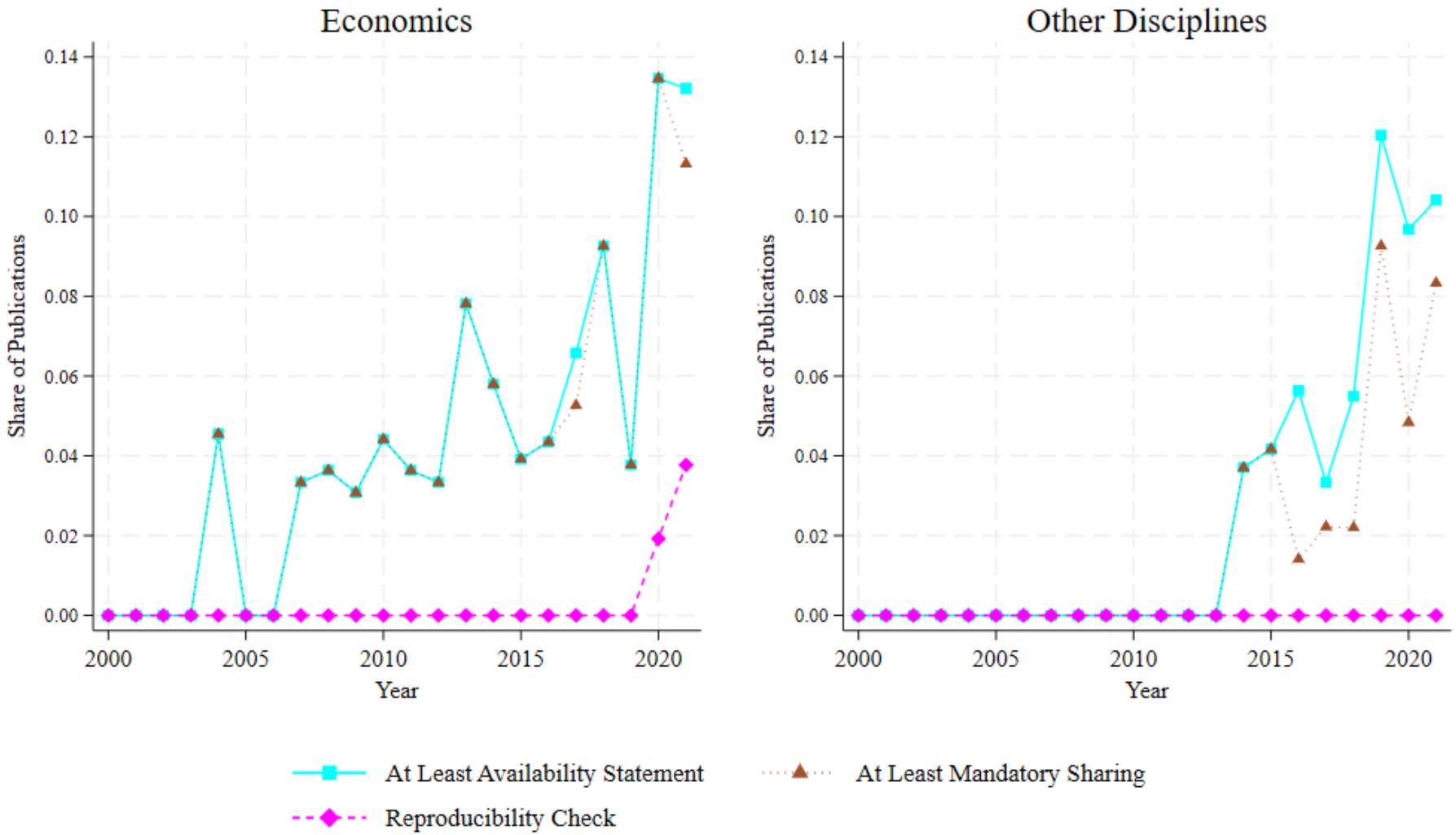
- **Uncertainty:** When specific policies were implemented.
- **Our approach:** Asked editors of journals with Level $\geqslant 1$ since when the policy is in place (n=98).
- **Response:** Received responses from 67 journals (68 %) on current status and from 62 journals (63 %) on when the policy was adopted
- **Extended search** Used Wayback Machine to find information for non-respondents

Journal policies over time

absolute numbers



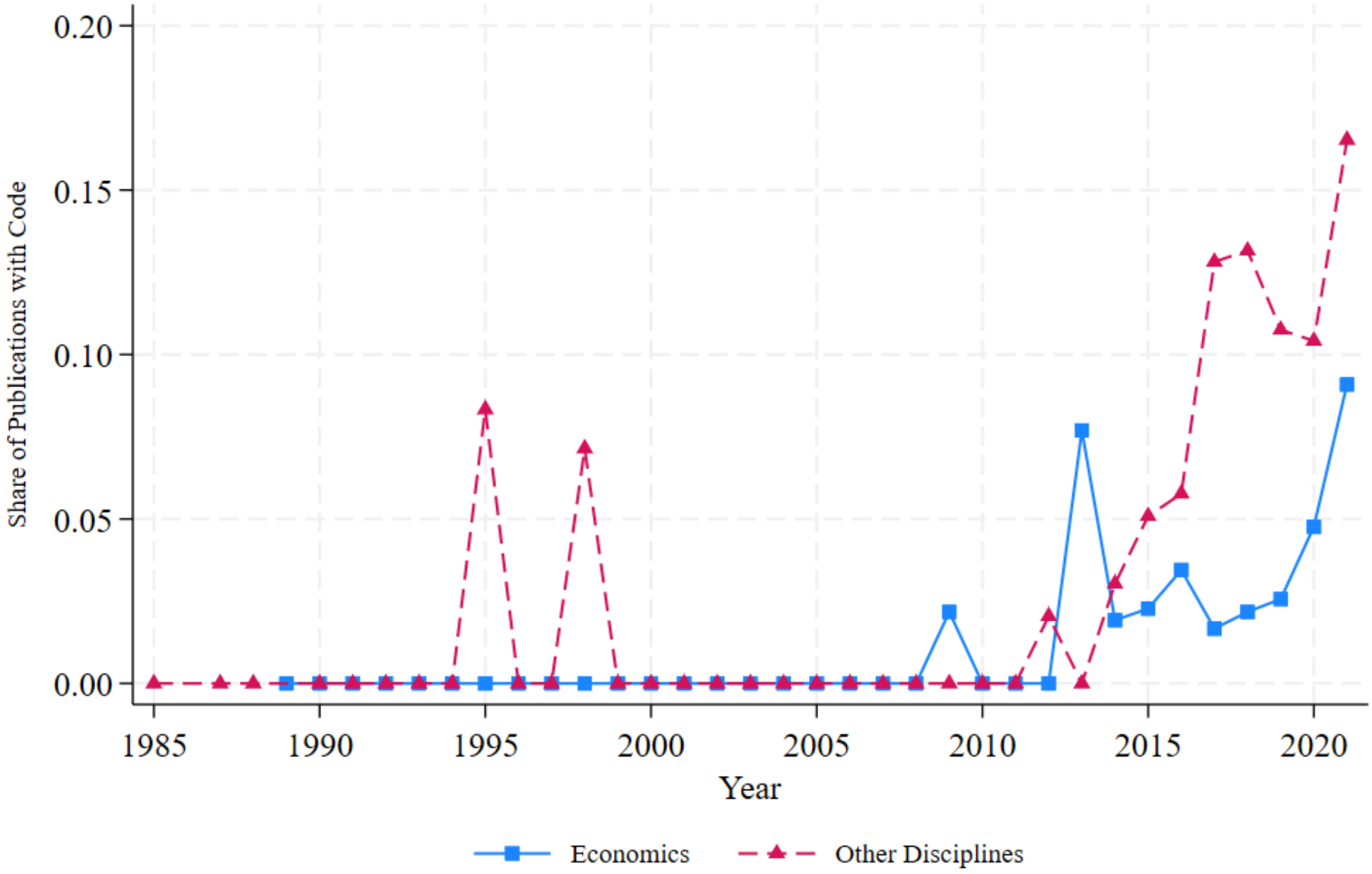
Articles covered by code-sharing policies over time



3. Awareness of researchers

- Negative examples of retracted studies
- Reproducibility/replicability crisis
  - Starting in the mid-2010s several comprehensive studies documented that many published articles cannot be replicated (e.g, in psychology, cancer research, experimental economics)
  - In economics, the debate around Herndon et al. (2013) and Reinhardt & Rogoff (2010) has also contributed to a growing awareness of reproducibility and code availability

Publicly available code, net of journal policies



Regression results: Predictors of code availability

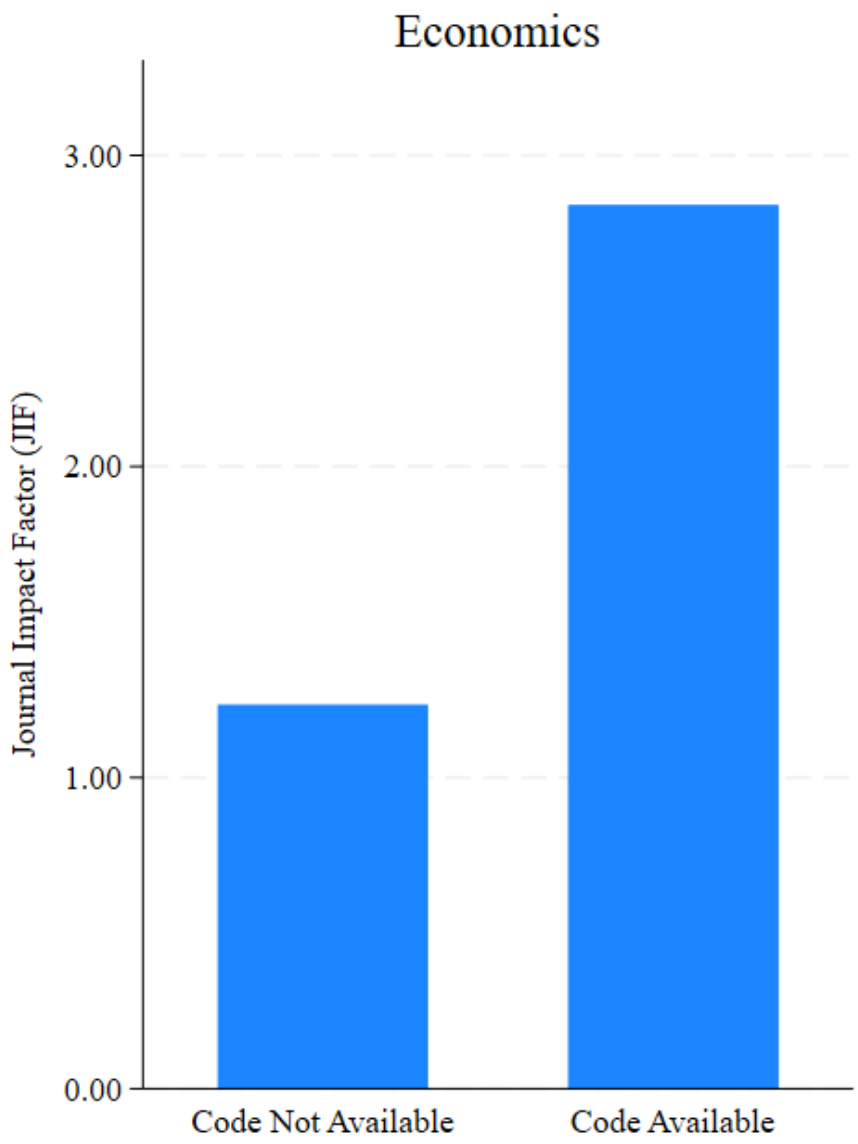
	(1)	(2)	(3)	(4)	(5)
Sociology	0.028*	0.025	0.027	0.043**	0.046**
	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
Psychology	0.057**	0.027	0.026	0.027	0.005
	(0.019)	(0.018)	(0.018)	(0.017)	(0.017)
Other Social Sciences	0.052***	0.029	0.030	0.040**	0.040**
	(0.016)	(0.015)	(0.016)	(0.014)	(0.015)
Health and Other Sciences	0.026	0.000	-0.001	-0.003	0.004
	(0.014)	(0.015)	(0.015)	(0.015)	(0.014)
English Article			0.000	-0.002	-0.002
			(0.014)	(0.014)	(0.014)
Single Author			-0.018	-0.011	-0.012
			(0.009)	(0.009)	(0.009)
Code Availability Statement				0.044	
				(0.076)	
Mandatory Code Sharing				0.366***	
				(0.053)	
Reproducibility Check				0.556*	
				(0.280)	
Publication Year FE	No	Yes	Yes	Yes	Yes
Only Level-0 Journals	No	No	No	No	Yes
F (4, N-1)	5.18***	1.96***	1.96***	4.35***	3.92***
N	2518	2518	2518	2450	2057

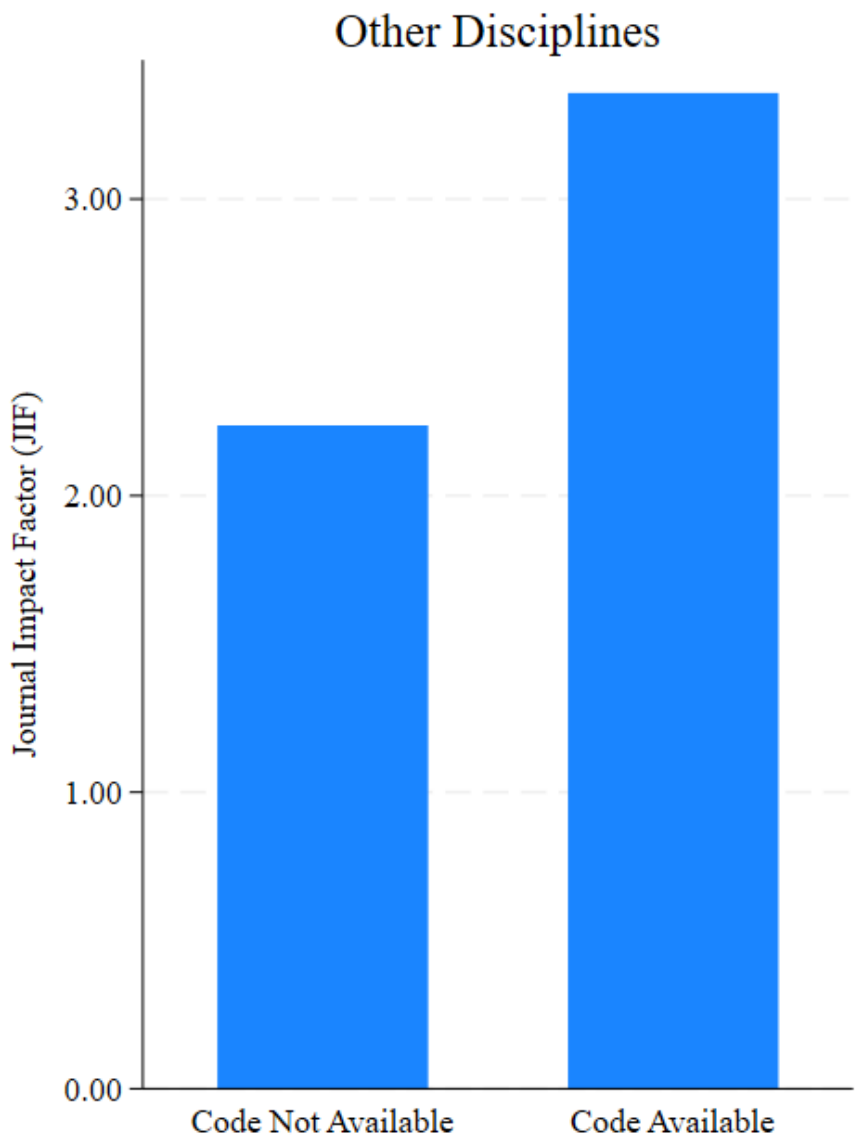
3. Results on replication code availability

3. Code availability and quality metrics

Code availability and 2-year journal impact factor

5-year



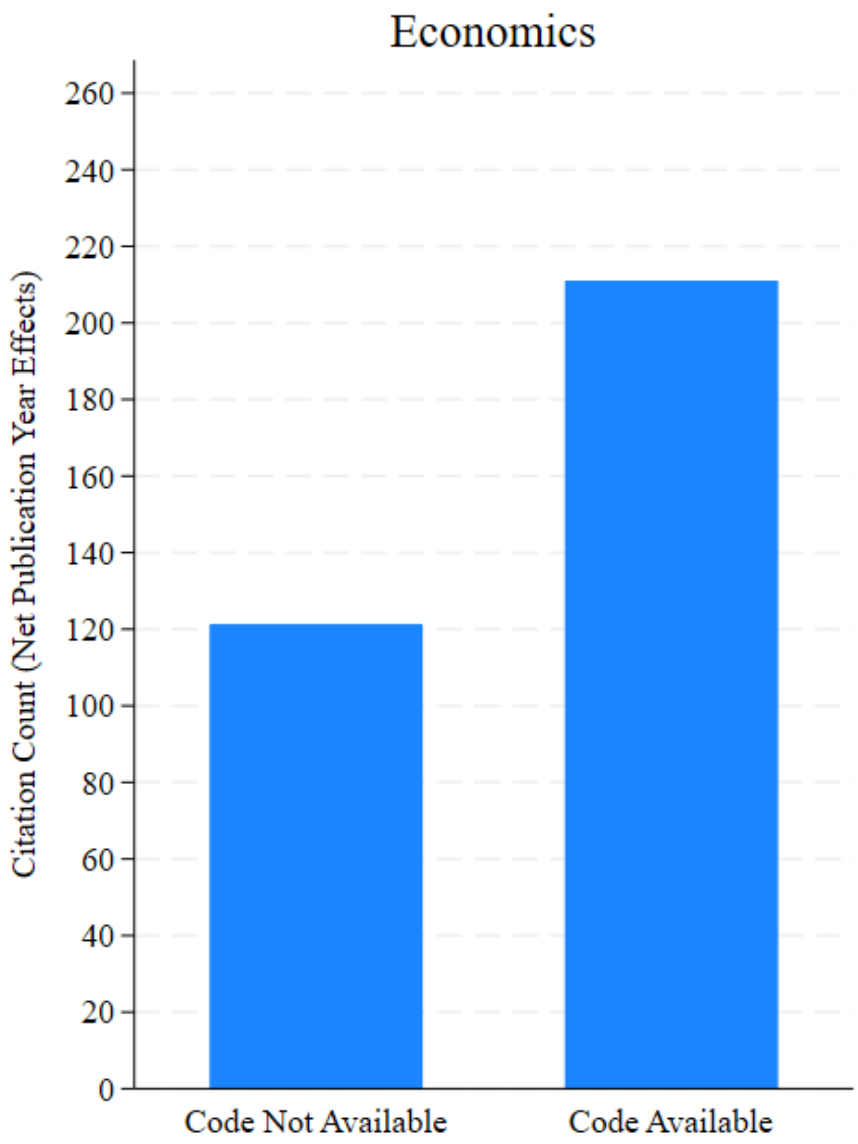


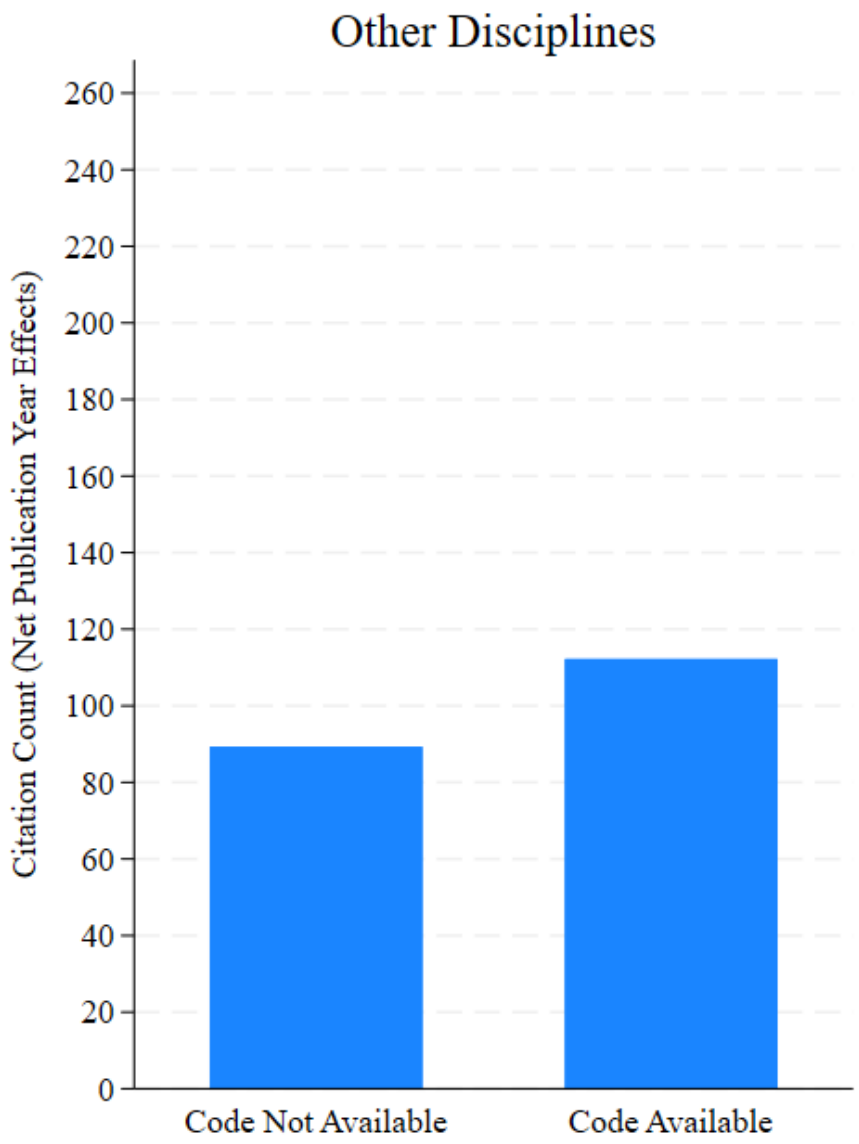
Multiple regression analysis for impact factor

Positive correlation between code availability and journal impact factor persists, also when controlling for

- Year of publication
- Principle discipline of the journal
- Language of the article
- Single-author
- Journal policies

Code availability and Google Scholar citations





Multiple regression analysis for citations

Positive correlation between code availability and citations persists, also when controlling for

- Year of publication
- Principle discipline of the journal
- Language of the article
- Single-author
- Journal policies
- Journal impact factor

3. Results on replication code availability

4. Robustness

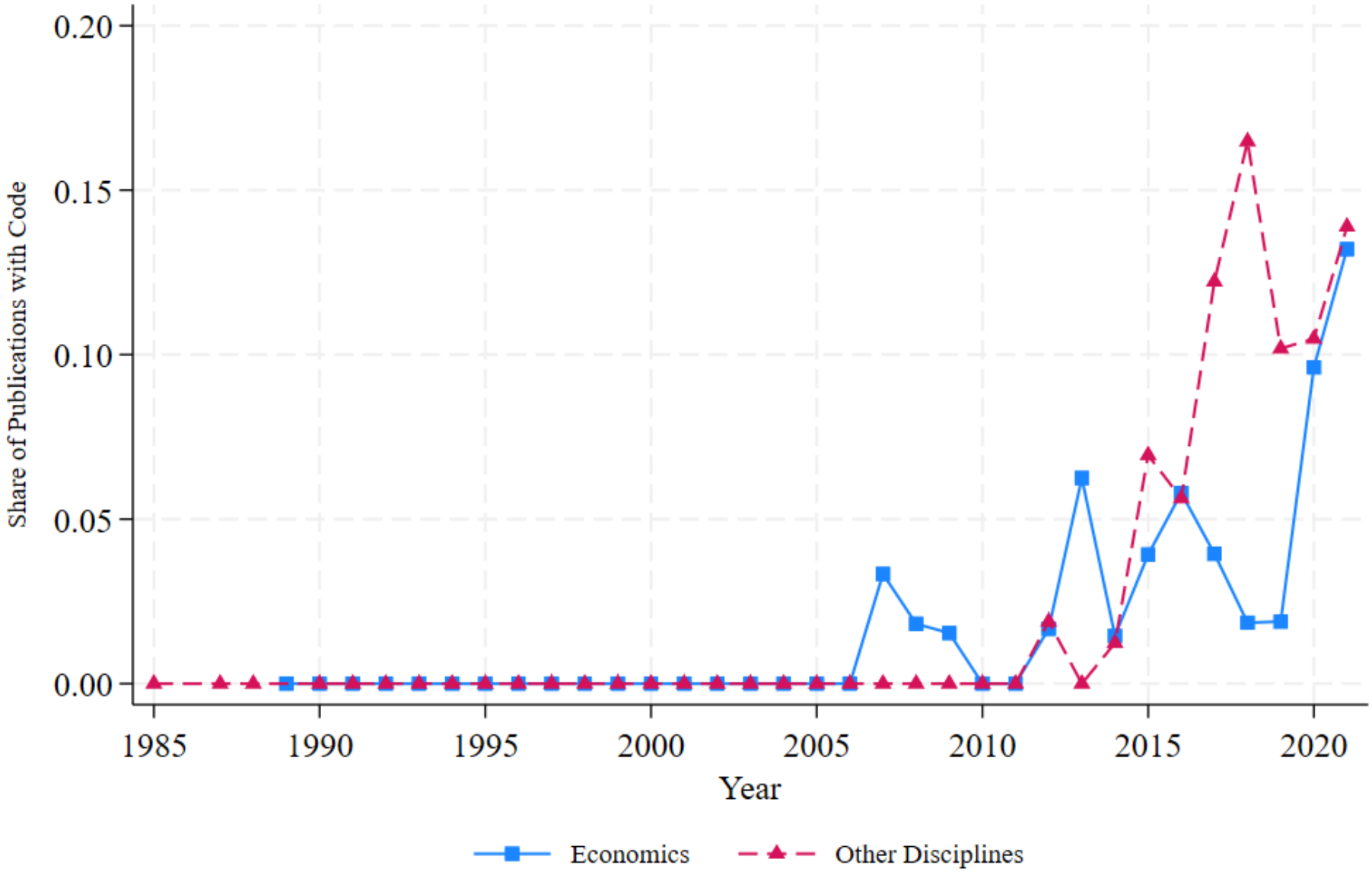
A stricter definition of code availability

- So far: code quality for reproducing results not taken into account
- Code availability does not ensure computational reproducibility
  - Due to e.g., missing data, non-transferability, unstable results, and differences in reproduced results.
- Reproducibility analysis of all 151 SOEP papers with provided code is beyond this paper's scope.
- Gertler et al. (2018) set a four-hour limit per paper for code runtime
- First step: Working with a stricter definition of code availability

Stricter definition: findings

- Stricter definition: Only raw SOEP data-loading code counted.
- 75% of code starts from raw data
- Using stricter measure, 4.5% provide code (113 in total).
- Evolution over time: Consistent increase in code availability.
- Quality measure: Limited improvement (75% in 2007-2011, 76.7% in 2012-2016, 75.7% in 2017-2021).
- Code provision increased, but quality remained relatively constant.

Publicly available code that loads raw SOEP data



4. Conclusion

Summary of findings

1. The share of SOEP studies with available code is relatively low.
2. The share of SOEP studies with code is increasing over time, both in economics and other disciplines.
3. Driven by a mixture of three factors
   - 3.1 Technological advances
   - 3.2 Top-down initiatives of journals
   - 3.3 Bottom-up initiatives of individual researchers
4. SOEP studies with available replication code are published in journals with higher impact factors.
5. SOEP studies with available replication code receive more citations.

   $\Rightarrow$ These correlations underscore the potential of code sharing as a quality signal.

Credibility revolution 2.0

- Leamer (1983: 37): "hardly anyone takes anyone else's data analysis seriously"
  - Title: "Let's take the con out of econometrics"

$\Rightarrow$ Angrist & Pischke (2010): this is no longer the case due to the "credibility revolution", the focus on credible research designs to identify causal effects

- "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics"

$\Rightarrow$ My view: credible research designs are not enough to counter the Leamer critique

- Researchers' degree of freedom: Researchers come to different conclusions when using the same data to analyze the same question (e.g., Huntington-Klein et al. 2021, Breznau et al. 2022)
- I have way more trust in research results if I can check all the details and easily run additional sensitivity checks

$\Rightarrow$ Need for a credibility revolution 2.0, where code (and data) sharing is the new normal

References

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2), 3-30.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44), e2203150119.

Cholewiak, S. A., Ipeirotis, P., Silva, V., & Kannawadi, A. (2021). SCHOLARLY: Simple access to Google Scholar authors and citation using Python. `https://github.com/scholarly-python-package/scholarly`

Gertler, P., S. Galiani, and M. Romero (2018). How to make replication the norm. *Nature* 554 (7693), 417–419.

Herndon, T., M. Ash, and R. Pollin (2013, 12). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics* 38 (2), 257–279.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco, G. Ekpe, T. Pugatch, M. Saavedra, and Y. Stopnitzky (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry* 59 (3), 944–960.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1), 31-43.

Nosek, Brian A., et al. (2015). Promoting an open research culture. *Science* 348.6242: 1422-1425.

Reinhart, C. M. and K. S. Rogoff (2010). Growth in a time of debt. *American Economic Review* 100 (2), 573–78.

Backup slides and appendix

Most frequent economics journals

Back

Economics

Labour Economics	79
Jahrbücher für Nationalökonomie und Statistik	62
Review of Income and Wealth	49
Economics Letters	48
Journal of Population Economics	44
Journal of Economic Behavior & Organisation	43
Health Economics	30
German Economic Review	29
Applied Economics	26
Empirical Economics	25
Economic Journal	24
Journal of Health Economics	22
European Economic Review	21
Journal of Human Resources	21

Most frequent sociology & psychology journals

Back

Sociology	
Kölner Zeitschrift fur Soziologie und Sozialpsychologie	98
European Sociological Review	94
Zeitschrift für Soziologie	67
Soziale Welt	17
Research in Social Stratification and Mobility	16
Psychology	
Social Indicators Research	76
Journal of Happiness Studies	31
Journal of Personality and Social Psychology	20
Psychology and Aging	20
Journal of Research in Personality	14

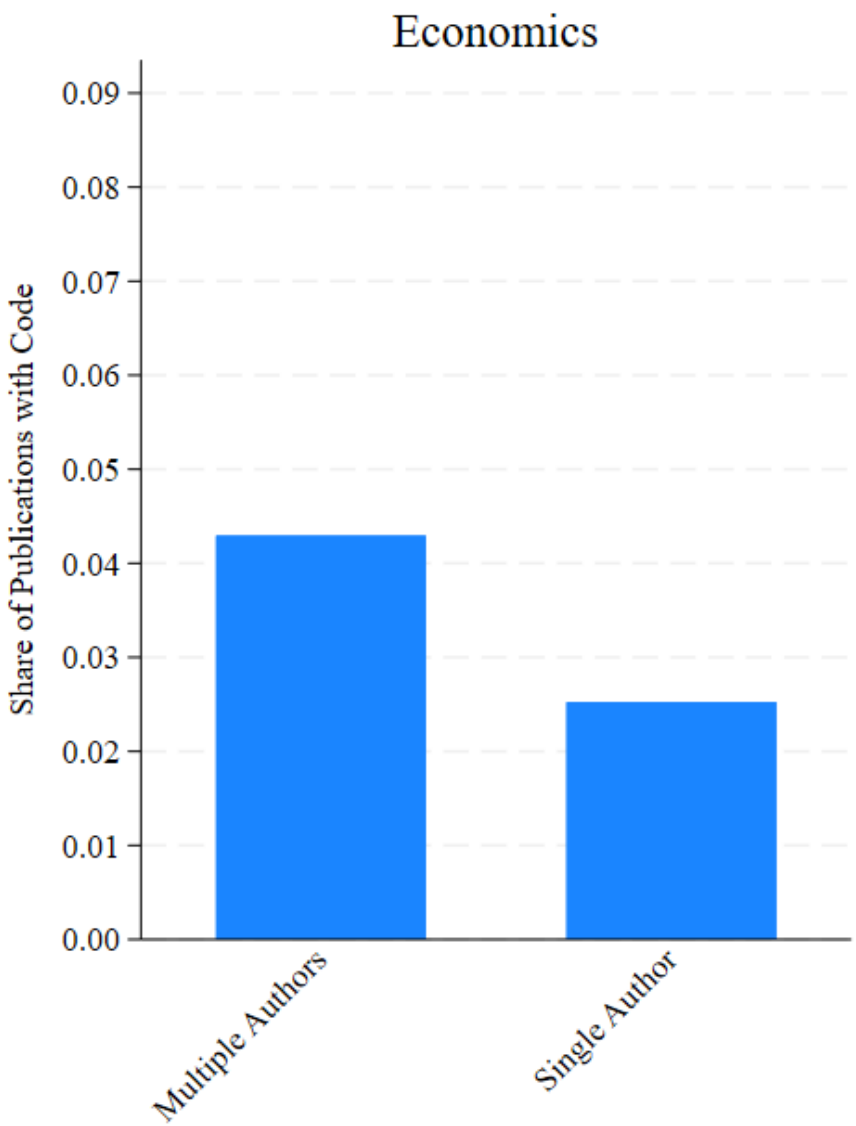
Most frequent journals: Other (Social) Sciences

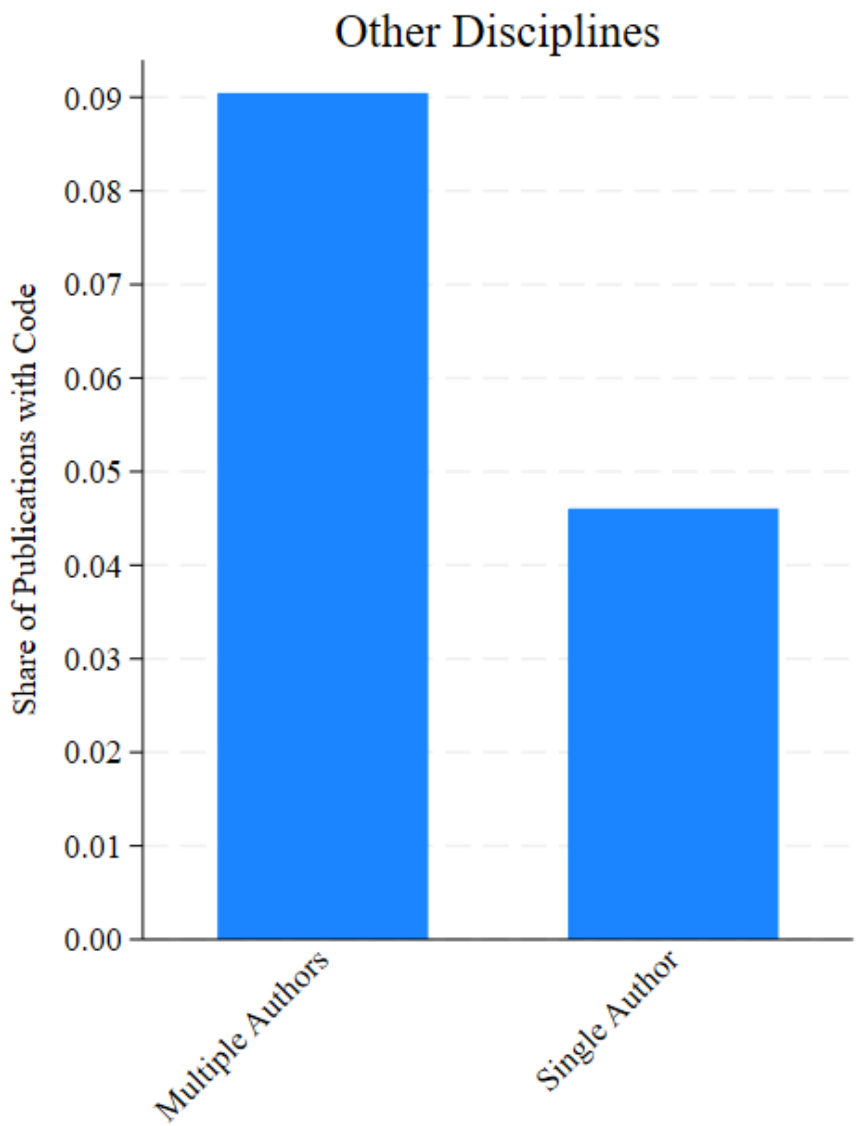
Back

Other social sciences	
Journal of Marriage and Family	29
Zeitschrift für Familienforschung	26
Journal of European Social Policy	18
Small Business Economics	18
Demography	17
Health & other sciences	
AStA-Advances in Statistical Analysis	35
PLOS ONE	24
Gesundheitswesen	19
Social Science & Medicine	19
Journals of Gerontology Series B-Psychological Sciences	13

Single author vs. multiple authors

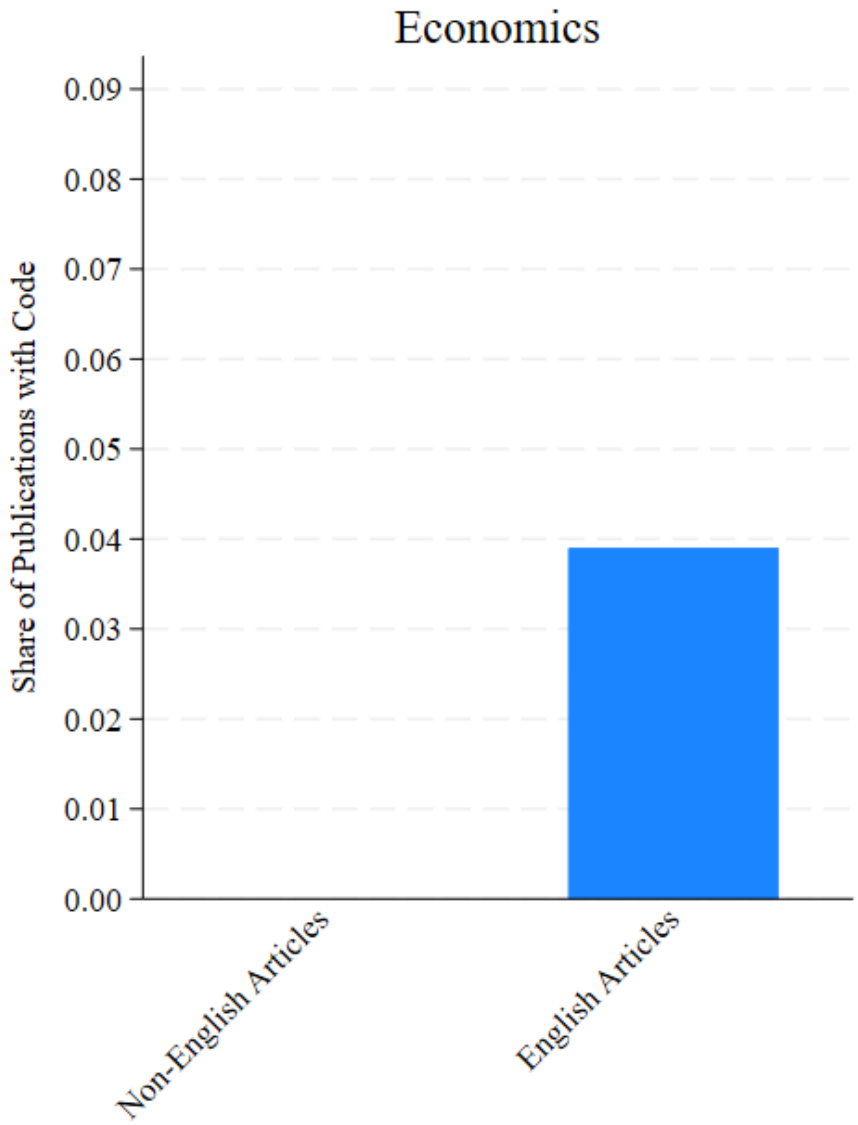
back

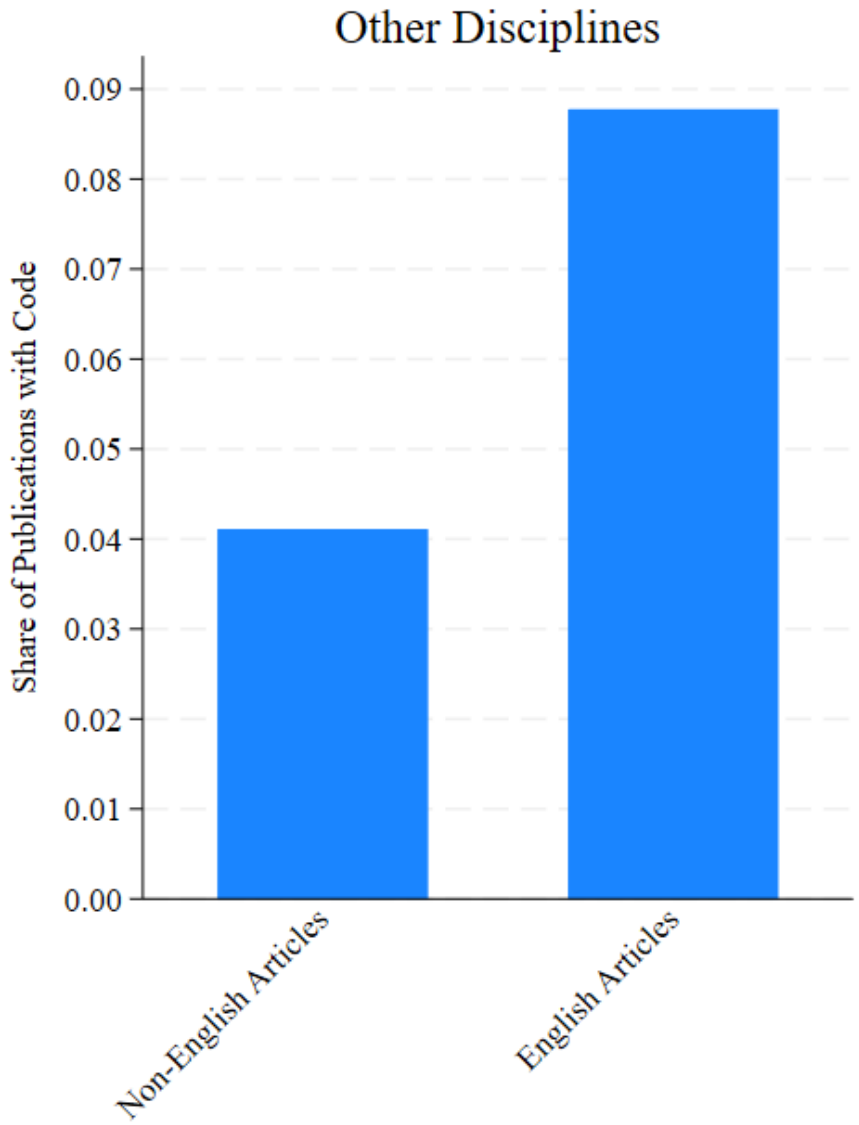




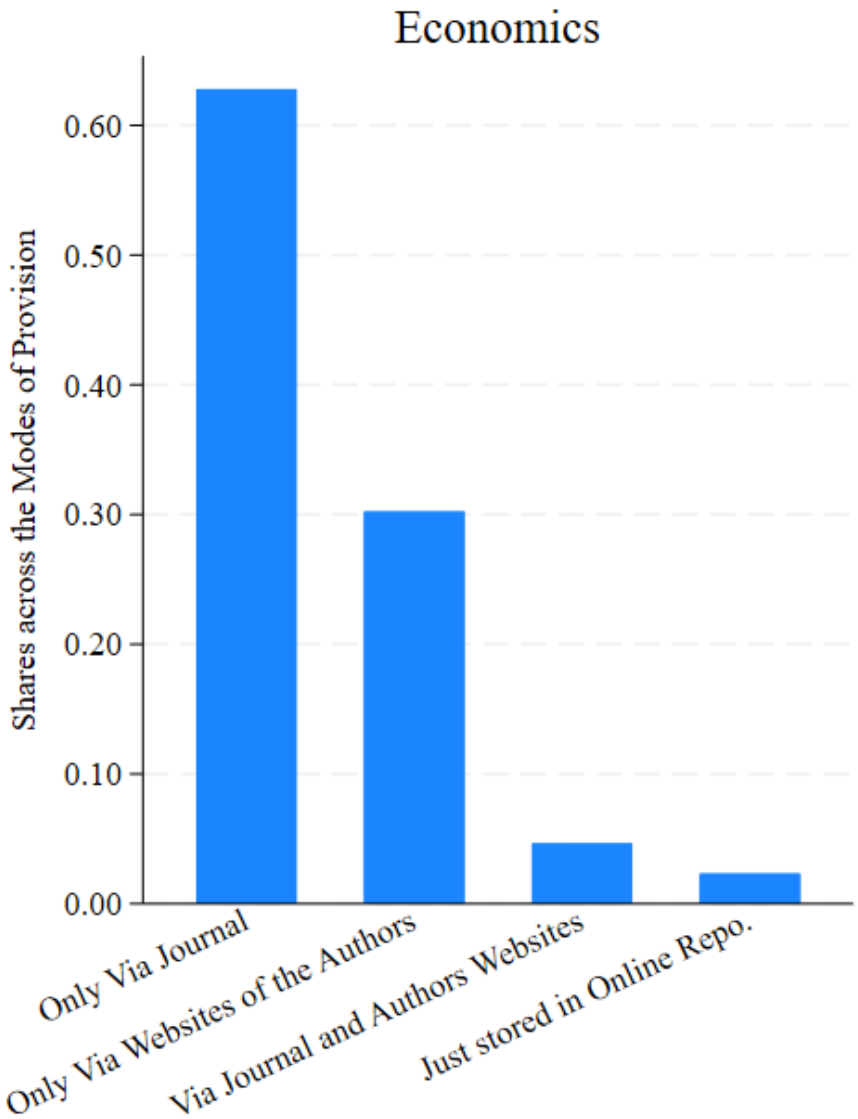
Code availability: Language of the article

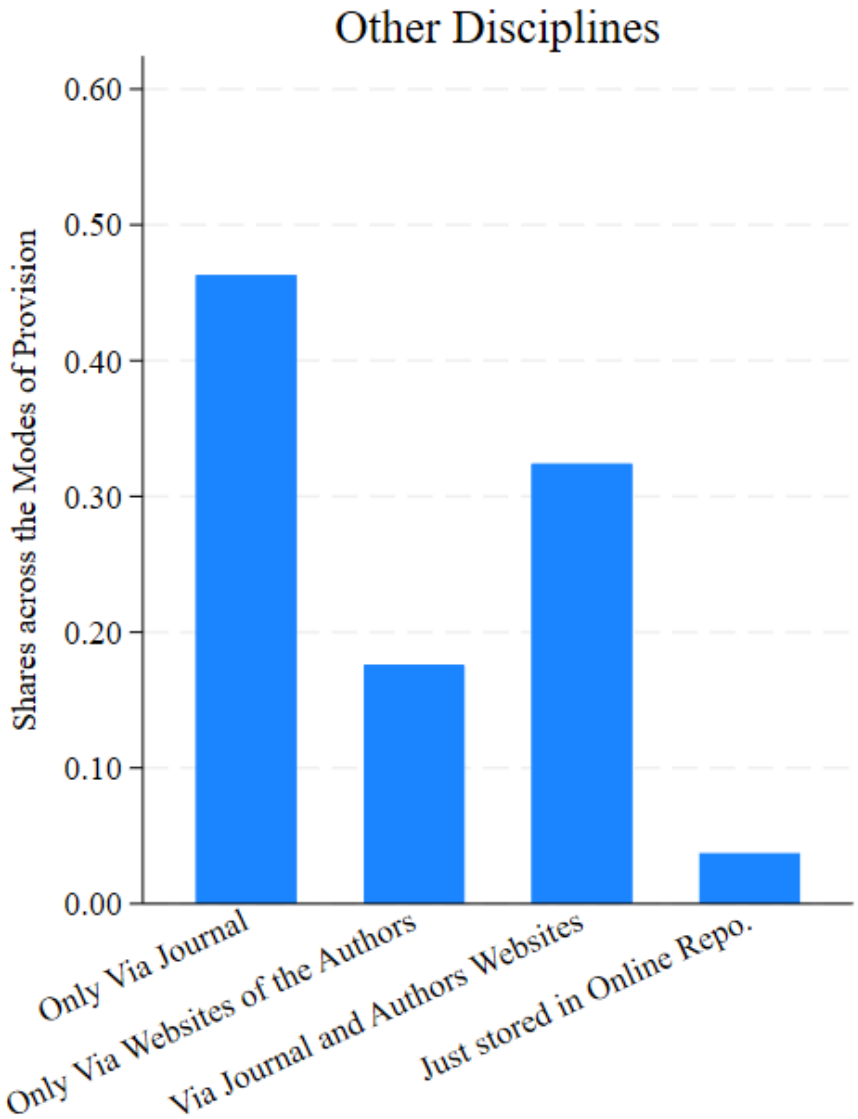
back





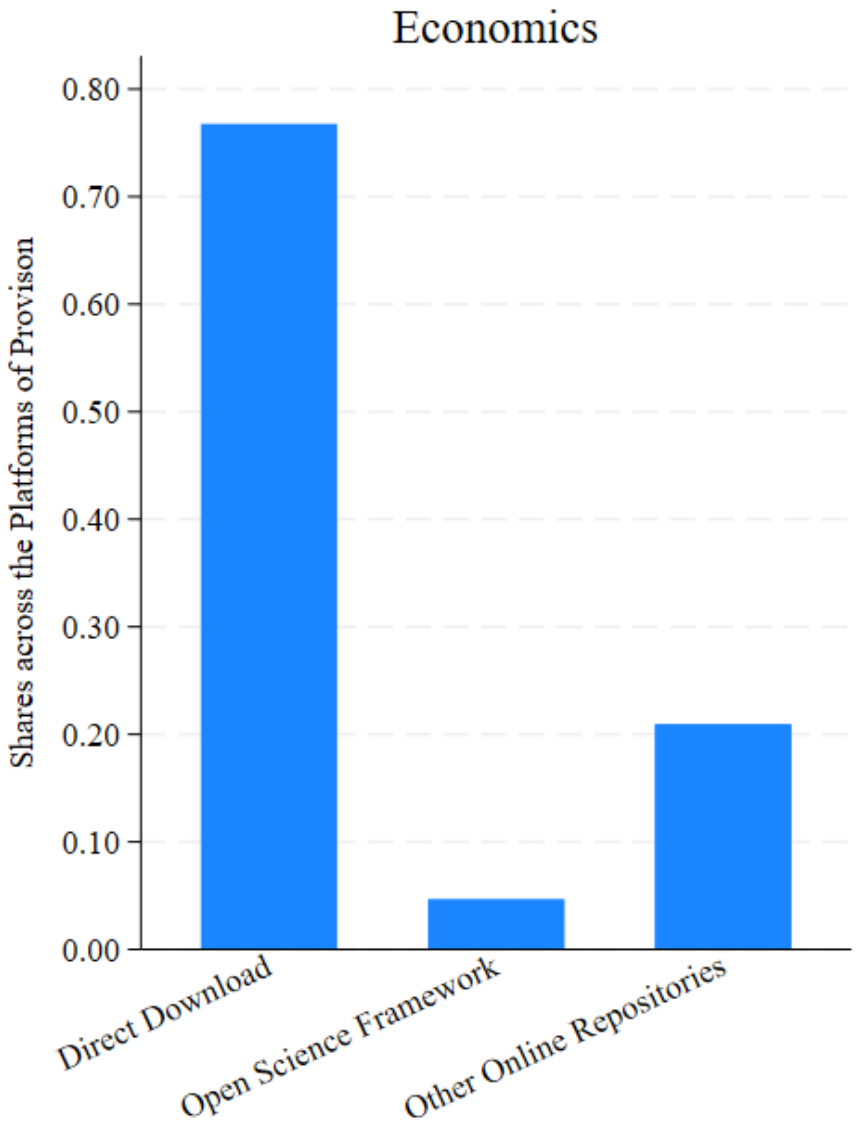
Mode of disclosure back

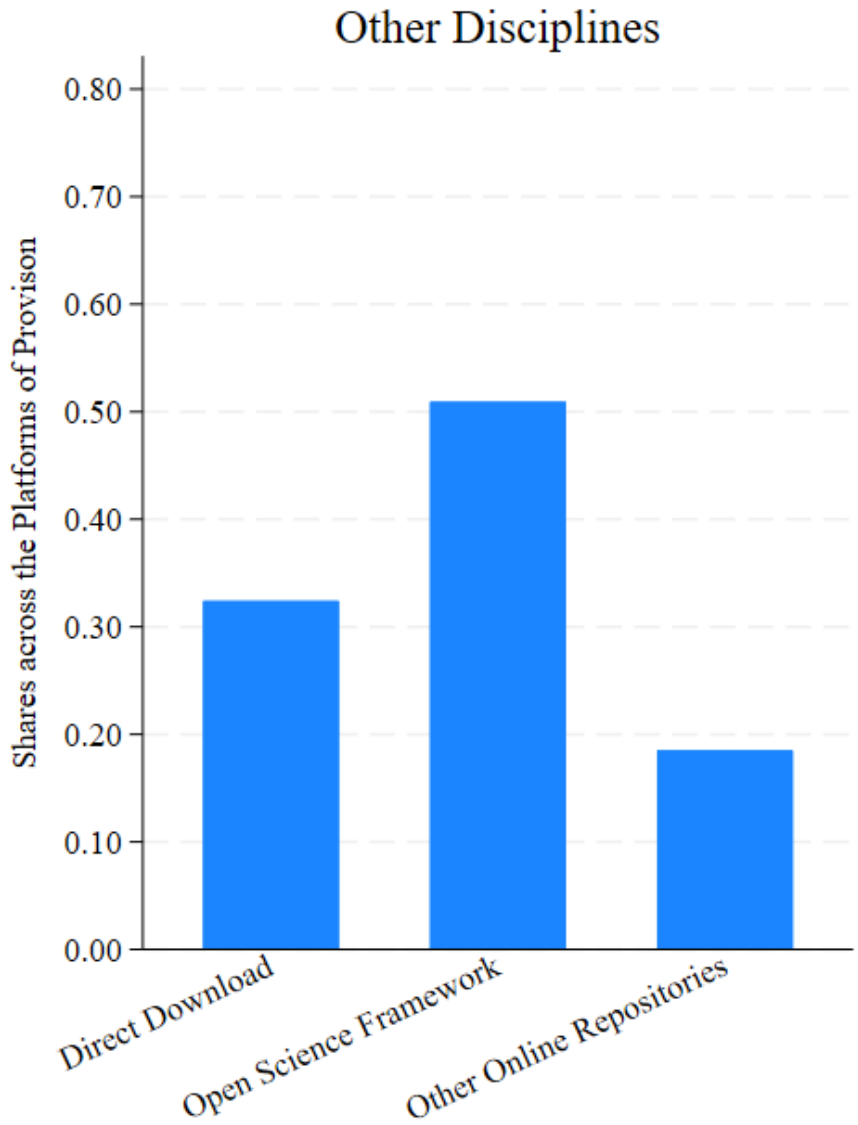




Storage location

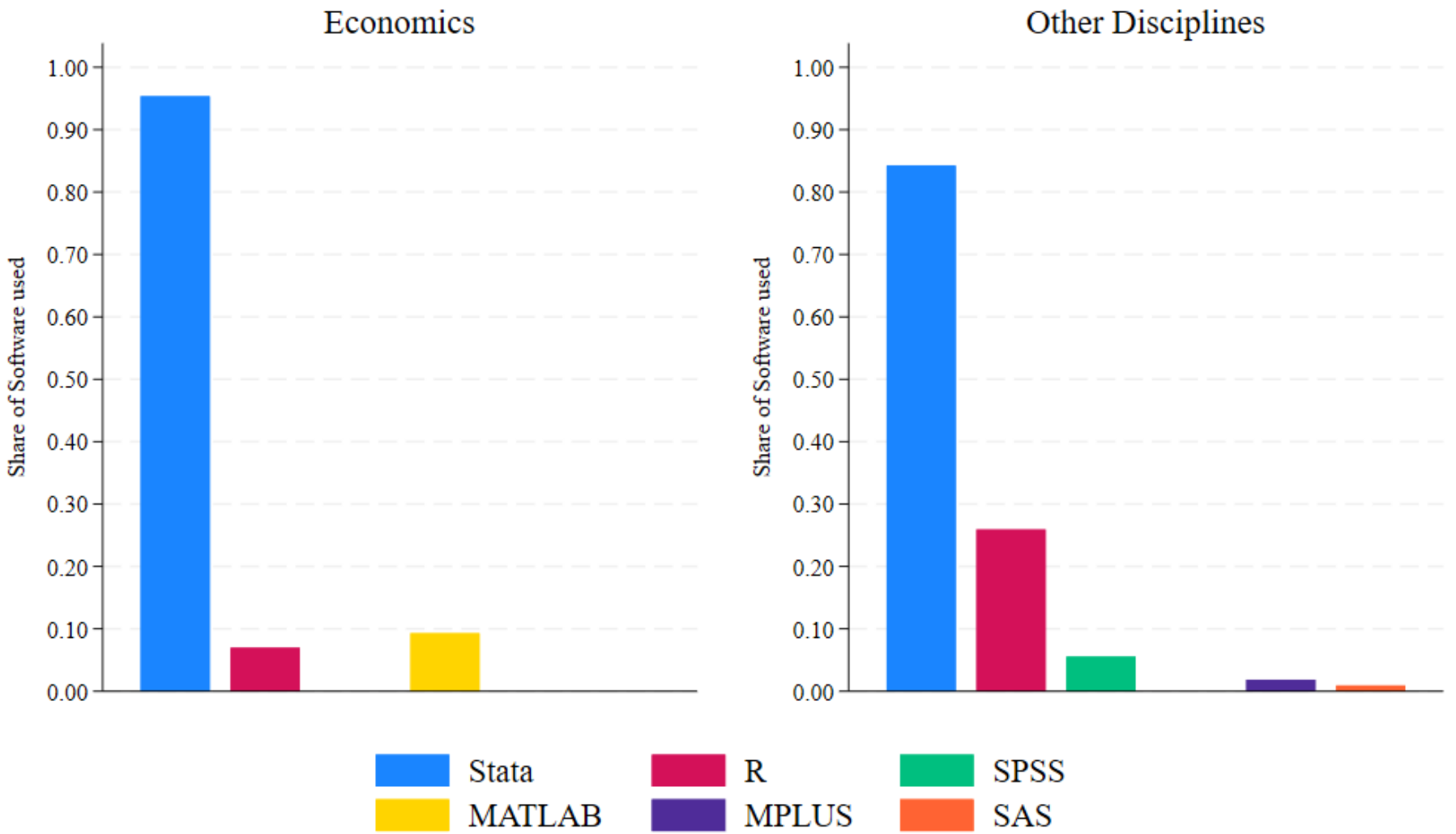
back





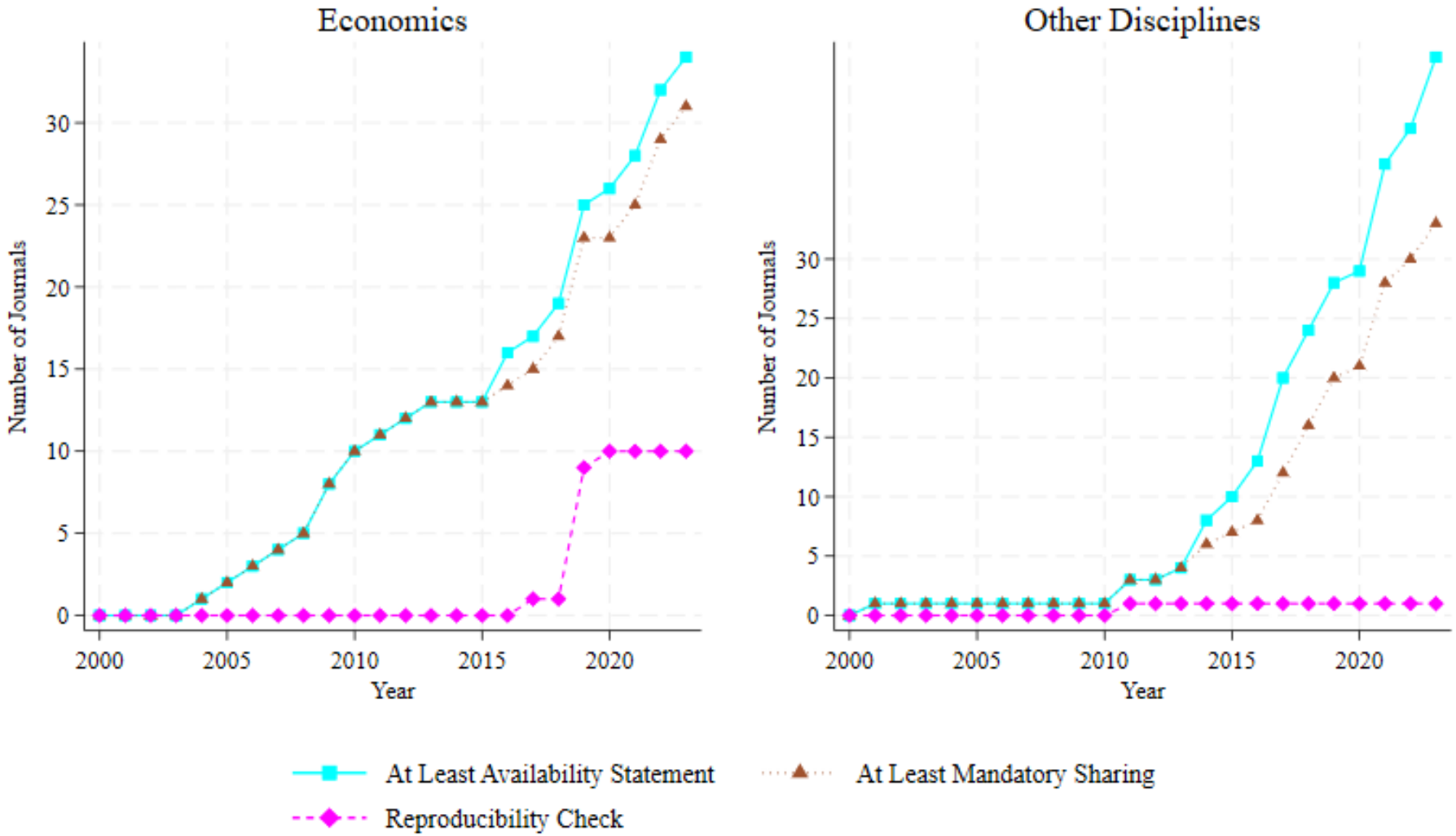
Shares across the software used

Back

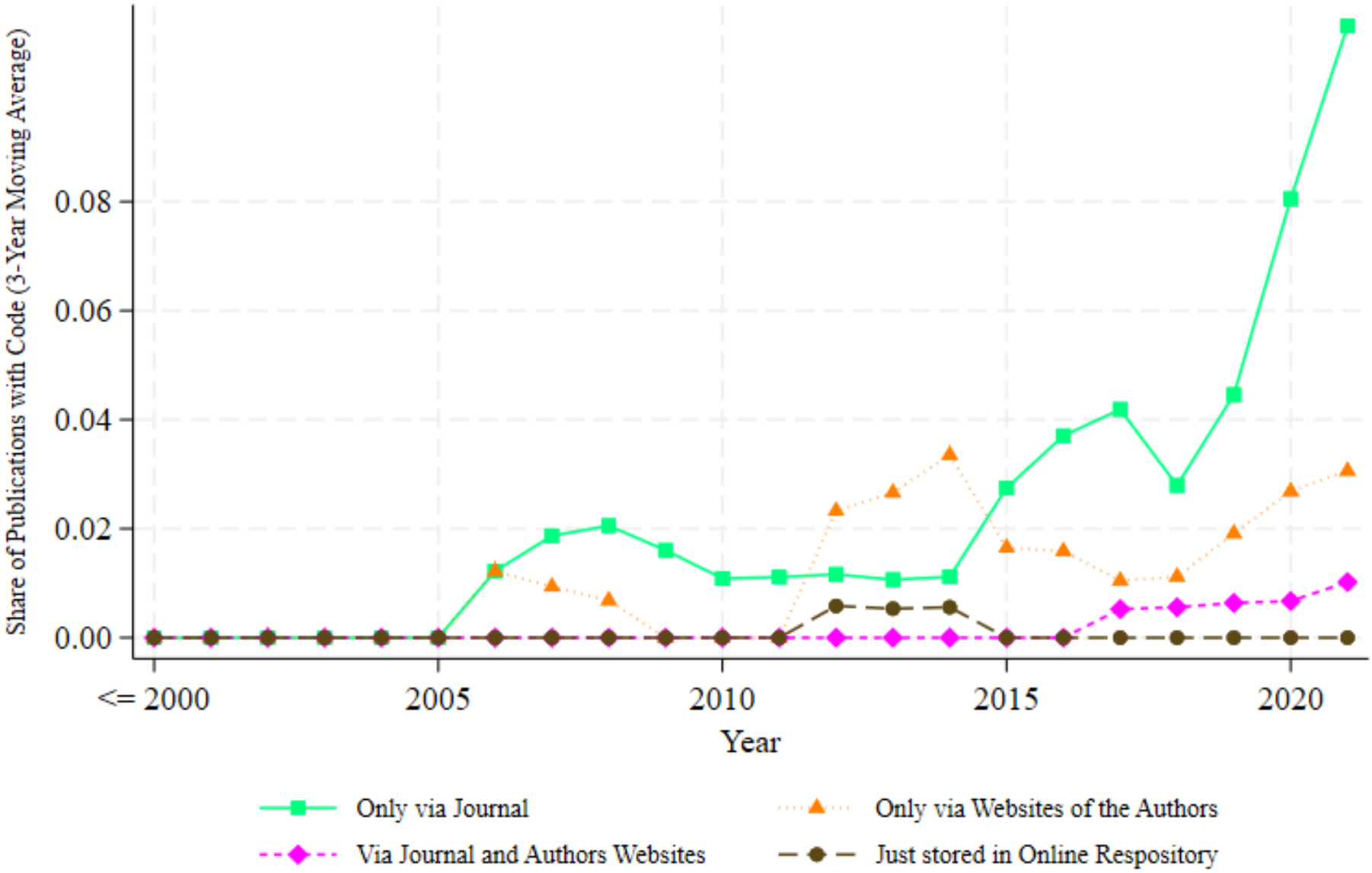


Journal policies over time

relative numbers

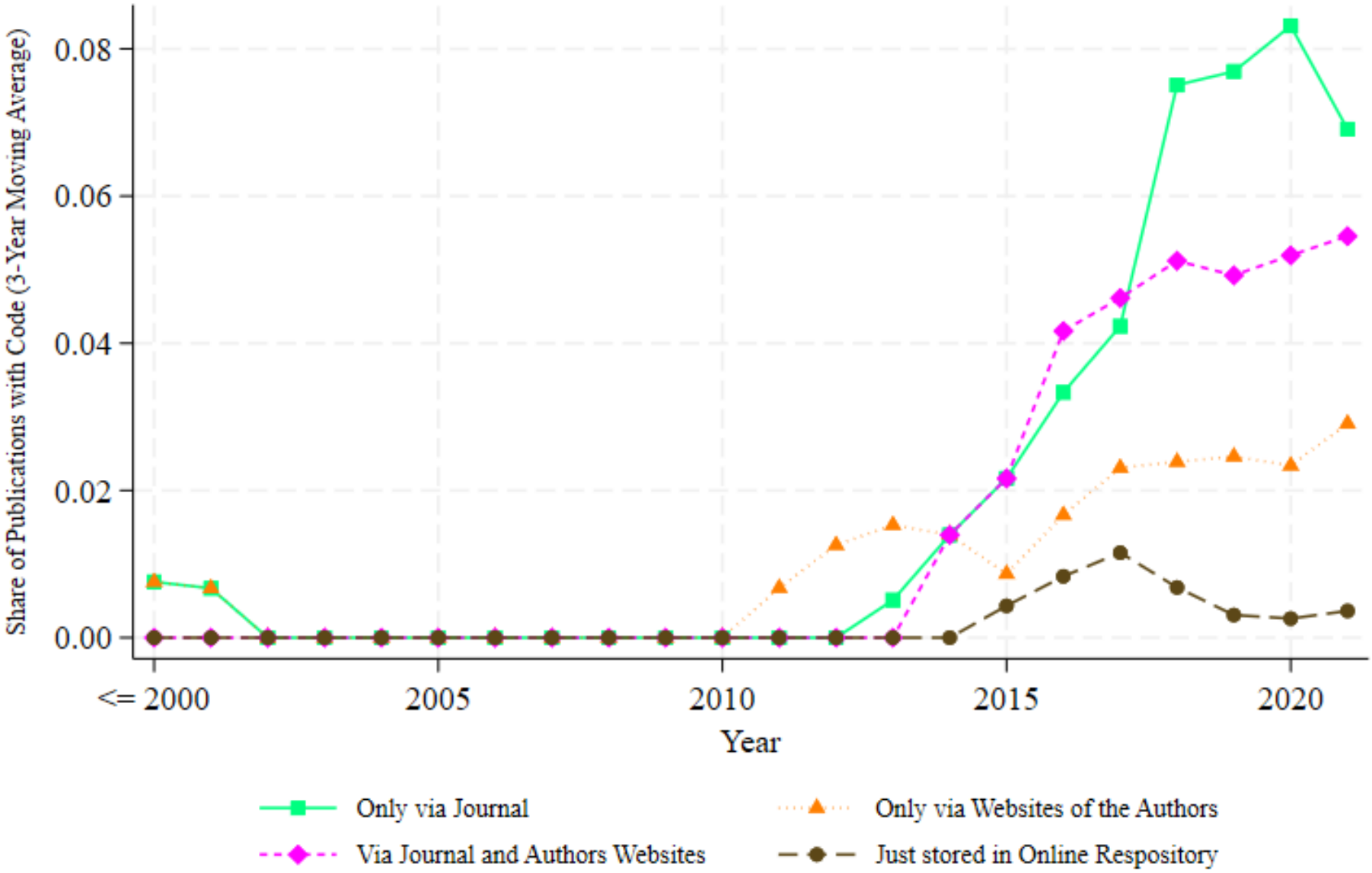


Mode of code provision (economics) back



Mode of code provision (other disciplines)

back



Code availability and 5-year journal impact factor

2-year