European Commission

JRC TECHNICAL REPORTS

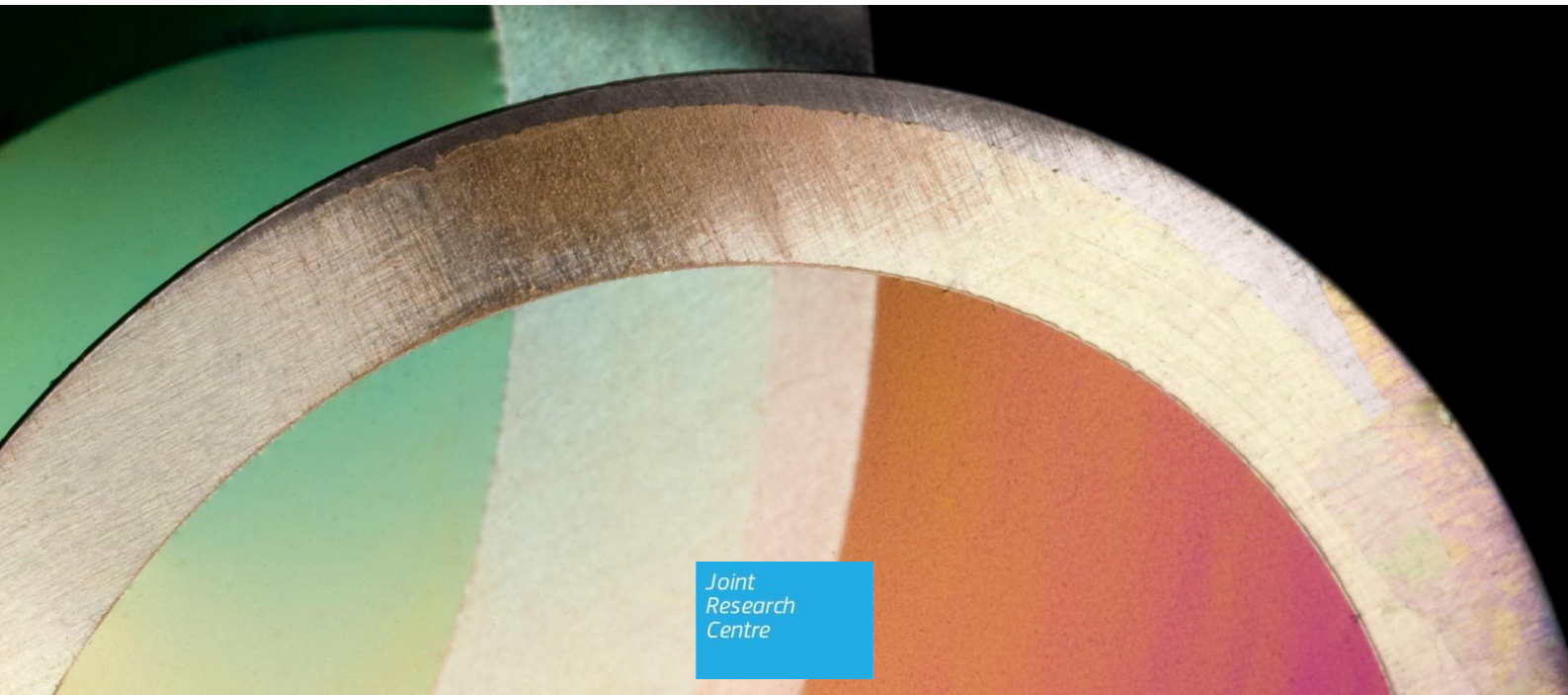
JRC Digital Economy Working Paper 2019-02

Does Data Disclosure Increase Citations? Empirical Evidence from a Natural Experiment in Leading Economics Journals

Mark J. McCabe[1,2] and Frank Mueller-Langer[3,4*]

[1] Questrom School of Business, Boston University, Boston, Massachusetts, USA
[2] SKEMA Business School, Université Côte d'Azur (GREDEG), Sophia Antipolis, France
[3] European Commission, Joint Research Center, Seville
[4] Max Planck Institute for Innovation and Competition, Munich

January 2019



This publication is a Working Paper by the Joint Research Centre, the European Commission's in-house science service. It aims to provide evidence-based scientific support to the European policy-making process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

Contact information
European Commission, Joint Research Centre
Address: Edificio Expo. c/Inca Garcilaso, 3. 41092 Seville (Spain)
E-mail: frank.muller-langer@ec.europa.eu
Tel.: +34 9544-88731

JRC Science Hub
https://ec.europa.eu/jrc

JRC115801

Seville, Spain: European Commission, 2019

How to cite: Mark McCabe and Frank Mueller-Langer (2019), Does Data Disclosure Increase Citations? Empirical Evidence from a Natural Experiment in Leading Economics Journals, Digital Economy Working Paper 2019-02, JRC Technical Reports.

Abstract

Does data disclosure have an impact on citations? Four leading economics journals introduced a data disclosure policy between 2004 and 2006. We use panel data consisting of 17,135 article citing-year observations from 1996 to 2015 for articles published in these journals. Empirical articles that did not disclose data (46% of the sample) serve as a control group. Evidence for a positive open data citation effect is weak (6% and not statistically significant). On the other hand, the citation impacts of publication are substantial and precisely estimated. Pure theory, hybrid and purely empirical articles enjoy citations benefits of 22%, 32% and 44%, respectively. Our pre- and post-publication citation data allow us to identify the citation effects of data disclosure and publication, while controlling for intrinsic article quality.

JEL Codes: L17, O33, C80, L59

Keywords: Data disclosure, diffusion of knowledge, natural experiment, panel data

* We thank Christian Zimmermann (Research Papers in Economics, RePEc) for providing us with citation data for the journal articles and working papers under study. We thank Penny Goldberg and Esther Duflo for providing us with information on the transition period and adoption of AER's mandatory data disclosure policy. We thank Luis Aguiar, Nestor Duch-Brown, Estrella Gomez-Herrera and Maciej Sobolewski for valuable comments. Michael Gerstenberger, Julian Hackinger, Benjamin Heisig and Christoph Winter provided excellent research assistance. The data was collected, the articles under study categorized, and first analyses performed when F.M.L. was senior research fellow at Max Planck.

1. Introduction

Open access to research data has attracted attention from economists (Dewald et al., 1986; Glandon, 2011; Hamermesh, 2007; McCullough et al., 2006; McCullough and Vinod, 2003) and policymakers (Burgelman et al., 2010; Doldirina et al., 2015; European Commission, 2012, 2016 & 2017; ESRC, 2019; NIH, 2003; NSF, 2014; OECD, 2007; US House of Representatives, 2007). For example, to overcome the gap between the large demand for replicable results and the low supply of data, the European Commission is encouraging access to research data generated by Horizon 2020 projects through the extended open research data (ORD) pilot. As the default setting, Horizon 2020 projects must deposit their research data in a research data repository.[1] Periodically, this subject sparks a fierce debate, whether due to data fraud (see Levelt Committee et al. (2012) for a discussion of the scandal surround the social psychologist Diederik Stapel) or mistakes, e.g. the coding errors made by two leading economists, Reinhart and Rogoff, in a pair of 2010 papers (see Herndon et al. (2014) and Bell et al. (2015)).[2]

Data disclosure is essential for the academic community and science policy. It improves the quality of research results, increases the efficiency of the academic system and pushes subsequent research (Anderson et al., 2008; Furman and Stern, 2011; McCullough et al., 2008; Nature, 2009). However, in applied economics, it is not common to voluntarily share research data (Andreoli-Versbach and Mueller-Langer, 2014; McCullough et al., 2006). The market for replication studies is limited and thus the incentives for researchers to write a replication study are low (Hamermesh, 1997; McCullough et al., 2006; Mirowski and Sklivas, 1991; Mueller-Langer et al., 2019). Hence, it may not be rational to invest

[1] The ORD pilot provides the option to opt out under certain conditions such as the protection of confidentiality.

[2] Carmen M. Reinhart and Kenneth S. Rogoff made their data available upon request to Thomas Herndon, Michael Ash and Robert Pollin. Reinhart and Rogoff (2013) provide an erratum to Reinhart and Rogoff (2010b). Reinhart et al. (2012) address some of the methodology and exclusion issues raised by Herndon et al. (2014). Finally, Bell et al. (2015) used Reinhart and Rogoff's data to re-examine the relationship between growth and debt in developed countries.

extra time and effort to create and store data in a manner that facilitates replication studies. Indeed, it may be rational for researchers not to disclose their data in order to delay or prevent attempts to replicate their results.[3] There are several reasons why authors may not choose to disclose their data (Costello, 2009; Dewald et al., 1986; Feigenbaum and Levy, 1993; Vlaeminck et al., 2013). First, compiling and documenting data is costly. Second, researchers might be concerned about others detecting mistakes or fraud (Nelson, 2009). Third, not disclosing data may protect a first-mover advantage (McCullough, 2009) that can be exploited in later publications (Anderson et al., 2008; Stephan, 1996).

However, (mandatory) data disclosure may also be beneficial for disclosing authors. For instance, it may serve as a means for high-quality researchers to signal the quality and robustness of their empirical results (Anderson et al., 2008; Dasgupta and David, 1994; Feigenbaum and Levy, 1993; Lacetera and Zirulia, 2011). Andreoli-Versbach and Mueller-Langer (2014) provide empirical evidence that better researchers are more likely to share their data voluntarily. Arguably, not controlling for the quality of articles would therefore lead to an upward bias in estimates of the citation impact of data disclosure.

Prior research provides evidence for a positive correlation between data disclosure and citations in various academic fields such as astrophysics (Dorch, 2012; Henneken and Accomazzi, 2011), geosciences (Sears, 2012), biomedical science (Piwowar et al., 2007; Piwowar and Vision, 2013), and peace studies (Gleditsch et al., 2003).

However, the methodologies used in these papers are not particularly robust: Dorch (2012), Henneken and Accomazzi (2011), and Sears (2012) simply compare the mean citation values of papers with and without data; the remaining papers employ regression methods using cross-section data. Each of these methods is particularly vulnerable to omitted variable bias due to unobserved article quality. Since

[3] Several studies tried to replicate empirical articles in economics (Camerer et al., 2016; Dewald et al., 1986; McCullough et al., 2006; McCullough et al., 2008) or other fields such as medicine (Alsheikh-Ali et al., 2011; Begley and Ellis, 2012). In the majority of the cases, it was not possible to replicate the articles under study.

article quality may be negatively or positively correlated with the willingness to share data (and therefore the number of citations), these studies may under/over-estimate the effect of data disclosure.

For these reasons, we believe the current literature does not provide a robust foundation for editorial and public policies vis-à-vis open data. The objective of our paper is to begin addressing this need by using unique panel data to identify the open data citation effect. In particular, we use citations to pre-prints available via RePEc as a proxy for the intrinsic quality of articles before they are published and receive the data-disclosure treatment. We benefit from a natural experiment in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, and the *Review of Economic Studies*. At different times during 2004-2006, each journal introduced a mandatory data disclosure policy (for authors of empirical articles), but not all authors complied; indeed, for 26.9% of the articles subject to this policy, no data was posted online. These non-complying journal articles contribute to the control group for the treatment group, i.e. empirical articles for which data is actually disclosed. Our main results are the following. Evidence for a positive open data citation effect is weak. The aggregate citation impact of mandatory data disclosure among empirical articles is 6% but is not statistically significant. On the other hand, the citation impacts of publication are substantial and precisely estimated. Pure theory and hybrid articles enjoy citations benefits of 22% and 32%, respectively. The benefits for purely empirical articles are 44%.

The remainder of the paper is organized as follows. Section 2 provides an overview of the related literature. In section 3, we describe our panel data. Section 4 outlines and discusses methodology. In section 5, we report the results of our empirical analysis. Section 6 considers the implications of these results for various stakeholders. Section 7 concludes.

2. Review of Related Literature

The literature studying the citation effect of data disclosure is relatively scarce. The few existing articles typically find a positive correlation between data disclosure and citations. However, the approaches and statistical methods that are used are not particularly robust.

One strand of the literature relies on descriptive statistics, i.e. a comparison of mean citation rates for articles with and without accessible data. For instance, Dorch (2012) studies citation advantages of articles with data-links by looking at articles in the *Astrophysical Journal* in the period from 2000 to 2010. Dorch (2012) finds that articles with links to data are cited 50% more often than articles without data-links. Henneken and Accomazzi (2011), who consider journals in the field of astronomy and astrophysics, obtain similar results (+20%). Although Henneken and Accomazzi (2011) attempt to only compare "similar" articles (they group them according to certain keywords), they still compare just the means of articles with and without a data link. Furthermore, a study by Sears (2012) investigates the effect of data availability on the citation rate in the field of geosciences. Sears (2012) reports that in the period from 1993 to 2010 articles published in the journal *Paleoceanography* have on average 35% more citations if the data is publicly available. Notably, the aforementioned studies do not examine a causal relationship between data disclosure and the number of citations because of their failure to control for article quality.

Other papers in this literature rely on regression analysis of cross-section data to identify the impact of data availability. Again, articles with and without available data are compared. Regression methods allow the authors to control for a variety of factors that are likely to influence article citation counts aside from data availability, e.g. the time since publication, the length of the article, the number of authors, etc. However, the effect of data availability is likely to be confounded with article quality. This and other omitted variables may bias the estimates reported in the cross-section literature.

For instance, Gleditsch et al. (2003) run a negative binomial regression on all articles published in the *Journal of Peace Research* from 1991 to 2001 and find that an article is *ceteris paribus* cited twice as

frequently if its data is available (in appendices, online or via a request to the author(s)). Their controls include authors' country and gender as well as type, age and length of the article. A mandatory data disclosure policy appears to have been introduced at some point during the period 1991-2001. According to the authors, as of 2003, this policy had "only been in place for a few years."

Piwowar et al. (2007) analyze a set of 85 cancer microarray clinical trial articles published during the period 1999 to 2003. About one-half of the articles made their data available in a public repository. Their linear regression estimates indicate that data availability is associated with a 69% increase in citations. They include three control variables (journal impact factor, date of publication and US authorship) but acknowledge that the "demonstrated association does not imply causation" (Piwowar et al., 2007, p. 3). They argue that high-quality trials may be more likely to share their data because of greater resources or more confidence in the results. Another obvious weakness of their study is the small sample size.

To address some of these shortcomings, Piwowar and Vision (2013) conduct a broader (cross-sectional) study by analyzing 10,555 studies that created gene expression microarray data; 25% of the papers deposited their data in public depositories. Controlling for more covariates[4] the positive effect of data availability is now quite modest. Studies that made their data available received 9% more citations (significant at the 5% level). Although adding more covariates may lower the potential bias associated with unobserved article quality, the inability to "difference out" this confounding factor remains.

Our paper also relates to a recent strand of literature on the increasing impact of empirical research within economics and beyond (Angrist et al., 2017; Angrist and Pischke, 2010; Hamermesh, 2013 & 2018). Angrist et al. (2017) document a rise in the influence of economic research on other disciplines. Differentiating between theoretical and empirical papers, they find that much of this rise can be attributed to growth in citations to empirical work. Similar to Angrist et al. (2017) we use an automated

[4] These covariates are: Date of publication, journal impact factor, open access status, number of authors, first and last author publication history, corresponding authors' country, institution citation history and study topic.

process to distinguish between theoretical and empirical economic articles. However, while they explore the "downstream" impact of article type (and field) on other disciplines, our attention is focused "upstream": on how features of scientific communication in economics, e.g. the transition from working paper to published paper, and the introduction of data disclosure policies, interact with article type in the generation of aggregate citations. Note that data disclosure may be one factor contributing to the growth in citations to empirical economic research.

We contribute to the aforementioned strands of literature by using a natural experiment that allows us to examine causal effects of data disclosure. Our panel data, consisting of pre- and post-publication citation data available via RePEc, allows us to control for the intrinsic quality of articles before they are published and receive the data-disclosure treatment. We are also able to isolate the citation effect of data disclosure from the citation effect of publication in top-5 economics journals.

3. Data and Variables

3.1. Data

We use panel data consisting of 9,895 article citing-year observations from 1996 to 2015.[5] We obtain this data from articles published between 2000 and 2012 in the four top-5 economics journals that introduced a mandatory data disclosure policy in the period under study: AER (2005, Vol. 95), Econometrica (2004, Vol. 72), JPE (2005, Vol. 113) and ReStud (2006, Vol. 73).[6]

We only consider articles for which RePEc pre-prints are available. We consider RePEc pre-prints published between 1996 (the starting year of RePEc coverage) and 2010 to have sufficient pre-publication citation information for journal articles. To this basic citation data and article information

[5] The number of observations is 17,135 for the augmented dataset including pure theory articles (see Section 5.2).

[6] Recently, QJE has also adopted a mandatory data disclosure policy. However, as of 14 July 2017, replication data and code is only available for three articles published in 2016 and 2017 in QJE's DataVerse.

we merge hand-collected information on data disclosure policies of journals and actual data disclosure at the article-level. Following Andreoli-Versbach and Mueller-Langer (2014), we searched the author guidelines and journals' websites for a description of data disclosure policies and information on the first issue when the policy was adopted to identify articles that are subject to a data availability policy. We contacted the journal editors to confirm our information on the implementation of data disclosure policies.

Our initial sample consists of 1,408 journal articles, 838 of which are empirical, and 570 purely theoretical. The 838 empirical articles are potentially subject to mandatory data disclosure; the purely theoretical articles are not. Although we exclude the purely theoretical articles from the initial analysis, we include them later to estimate the corresponding publication effects. For the 838 empirical articles, we distinguish between articles that do not contain a theoretical model ("purely empirical") and those that combine theory and empirical work ("hybrid articles"). We use a semi-automated process based on theory- and empirical-related words that we retrieved from the article PDFs. This method classifies the articles as purely theoretical, purely empirical or hybrid (we describe the classification process in more detail in section 3.2). Regarding the 838 empirical journal articles, 592 were published after the corresponding journal introduced a data disclosure policy.[7] However, data is *currently* available for only 448 of these 592 articles; these 448 articles constitute our data disclosure treatment group. The remaining 390 articles for which no data is available act as controls (246 of these articles were published prior to the introduction of a data policy; 144 are subject to a data policy but are non-compliant).

Data availability today *may* imply compliance from the date of publication. To test this assumption we randomly selected 51 articles from three of our 4 journals (the AER, Econometrica and ReStud, which account for 410 of the 448 data-compliant articles). We then used the Internet Archive to confirm that

[7] Data is available for six AER articles published *prior* to the official introduction of the AER's data policy. These articles are included in the reported results. Excluding the 6 articles has no material impact on the results.

datasets were indeed available at the time of publication. Extrapolating from this result, we assume that all 410 articles were data-compliant upon publication.

Unfortunately, we could not use this approach for the remaining 38 articles published in the JPE; the JPE website blocks (Internet Archive) robots.

3.2. Classifying Articles

Using an automated classification process, we determine whether an article is of purely theoretical nature, purely empirical nature or has elements of both styles, i.e. "hybrid" articles.[8] We classify an article as purely empirical if the author(s) generated their main contribution from the use of data and empirical estimation techniques. We classify an article as purely theoretical if the authors derive their results exclusively via mathematical and axiomatic deductions, i.e., they do not use any data to arrive at their results. We classify an article as hybrid if it contains elements of both styles. We define an article as "empirical" if it is either purely empirical or hybrid. Notably, we will eventually use the automated classification process to distinguish between purely theoretical and empirical articles. In this case, the automated process achieves a rate of correct style prediction of 96% (see Appendix A). However, in order to distinguish between hybrid and purely empirical articles *within* the set of empirical articles we relied on the more precise manual classification procedure.

Generally, the idea behind our automated classification process is to set up a relationship between a set of style-specific words within an article and its corresponding style. The dependency of article style on the set of specific words within an article is estimated by probit regressions. The underlying idea is that the word composition within each article will characterize the article regarding its style. From the coefficients, which we receive by regressing the (manually) pre-determined article styles on their

[8] See also Angrist et al. (2017) who use machine learning techniques to classify articles as empirical, theoretical and econometrics.

corresponding word counts, we predict the article styles of the rest of the sample. While Appendix A provides a detailed description of the style classification process, here we briefly describe it.

The initial selection of theory-style specific and empirical-style specific words is derived from the manual classification of a random sample of 200 articles. In a final step, we then use another random sample of 100 articles (henceforth, test sample) to assess the accuracy of the classification process. To do that, we compare the predicted styles of the test sample with the manually classified styles of the test sample. When we consider two styles only, i.e., purely theoretical vs. empirical, we achieve a rate of correct prediction of 96%. When we consider three styles, i.e., purely theoretical, purely empirical and hybrid, we achieve a rate of correct prediction of 86%. For the prediction process with three styles, we examine the 14 articles where manual and automated classification lead to different styles in order to determine the source of the prediction errors. The automated process classified nine purely empirical papers as hybrid papers, three hybrid papers as purely empirical papers, and two purely theoretical papers as hybrid papers. Due to these assignment errors, we manually checked all empirical *and* hybrid articles. This resulted in the re-classification of 68 purely empirical papers as hybrid papers, 80 hybrid papers as purely theoretical papers, and 25 hybrid papers as purely empirical papers.[9] Finally, while our

[9] This manual exercise also revealed some potentially relevant differences across empirical papers in our sample. Although most empirical papers in our sample test hypotheses by applying econometric techniques to some type of real data, 72 hybrid papers deviate from this template and employ simulation techniques, e.g. Monte Carlo methods, or "quantitative" calibration. To determine whether the estimated publication and data disclosure effects for hybrid papers (reported in Table 3) differ across econometric and simulation-based papers, we specified a pair of publication and data disclosure dummies for each of these hybrid paper types. The results are starkly different: econometric-based papers are associated with large publication citation effects (30-40% range), but exhibit no citation impact from data disclosure. These results are reversed for the simulation-type hybrid papers: large data disclosure effects (30-40% range) but no citation effect from publication. Further examination of this simulation "anomaly" suggests that the data disclosure effects are associated solely with the calibration papers. Whether this anomaly is real or a result of small size is unclear.

approach was designed independently from the approach adopted by Angrist et al. (2017),[10] there are interesting parallels. For instance, four articles reported in Table A2 of Angrist et al. (2017, p. 53) are also included in our sample. We compare the respective classification results. First, Angrist et al. (2017, p. 19) classify "Christiano, Eichenbaum and Evans (2005) as empirical, even though they combine theory with empirical work". The combination of theory and empirical work constitutes our definition of a hybrid article. Our automated approach classified Christiano, Eichenbaum and Evans (2005) as hybrid. Second, Acemoglu, Johnson and Robinson (2001) is classified as "empirical" in Angrist et al. (2017) while our process classifies this article as (purely) empirical. Third, Kling, Liebman and Katz (2007) is classified as "empirical" in Angrist et al. (2017) while our process classifies this article as (purely) empirical. Fourth, Melitz (2003) is classified as "theoretical" in Angrist et al. (2017) while our process classifies this article as "purely theoretical" (and consequently it is excluded from our main analysis). These comparisons provide additional confidence that the results of our classification process are correct.

4. Empirical Methodology

4.1. Natural Experiment

Recently, several major economics journals introduced mandatory data disclosure policies, which require authors to share their data prior to publication. The leading example is the mandatory data disclosure policy of the AER, which other journals adopted.[11] For instance, AER (2019) states the following:

[10] Angrist et al. (2017)'s machine learning algorithm uses an updated training sample that was original used in Ellison (2002) to classify empirical articles if they use real-world data and report econometric estimates. Notably, Angrist et. al.'s empirical style overlaps with our definition, so it includes hybrid articles.

[11] Notably, the first economics journal that introduced a mandatory data disclosure policy is the Federal Reserve Bank of St. Louis Review (1993, 75(1)).

"*It is the policy of the American Economic Association to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the AEA website.*"

Identifying the citation impact of these data disclosure policies poses several challenges: first, we need to control for article quality so that comparisons of our treatment and control articles (those with and without posted data, respectively) are not subject to an omitted variable critique. Second, citation effects are likely to arise when a working paper is published. Third, both data disclosure and publication citation effects may vary based on the type of empirical articles considered: hybrid or purely empirical.

Clearly, to estimate publication and post-publication data disclosure effects, we need to observe citations to articles before and after publication. This allows us to control for article quality (by specifying article fixed effects). To distinguish further between the citation impacts of publication and data disclosure (which occur simultaneously for articles in journals that have adopted a data policy), we need to check if compliance is partial, and then specify a data dummy that equals one for compliant articles.

Our panel data consist of citations to articles *before* and *after* publication in several journals.[12] This allows us to control for intrinsic article quality by specifying article fixed effects. Due to data policy non-compliance, we are able to identify separate publication and data disclosure effects. We also test whether the citation effects of publication and data availability vary by article type.[13]

[12] This approach is possible as RePEc unambiguously tracks citations to a given paper during its creation and publication cycle, i.e. starting from the date when the first pre-print is posted on RePEc and continuing after publication.

[13] To estimate publication and data-disclosure effects, in the aggregate, and across article types, we rely on articles that were published before and after our journals' data policies were introduced. If we restrict the data only to articles subject to a data

4.2. Panel Count Data Specification

Following McCabe and Snyder (2015), we use Wooldridge's (1999) Poisson quasi-maximum likelihood (PQML) estimator[14] to account for the count data nature of citations in our panel data setting with the following conditional mean,

$$E(cite\_count_{it}|\alpha_i, \delta_{kt}, x_{it}) = \exp(\alpha_i + \delta_{kt} + x_{it}\beta), \quad (1)$$

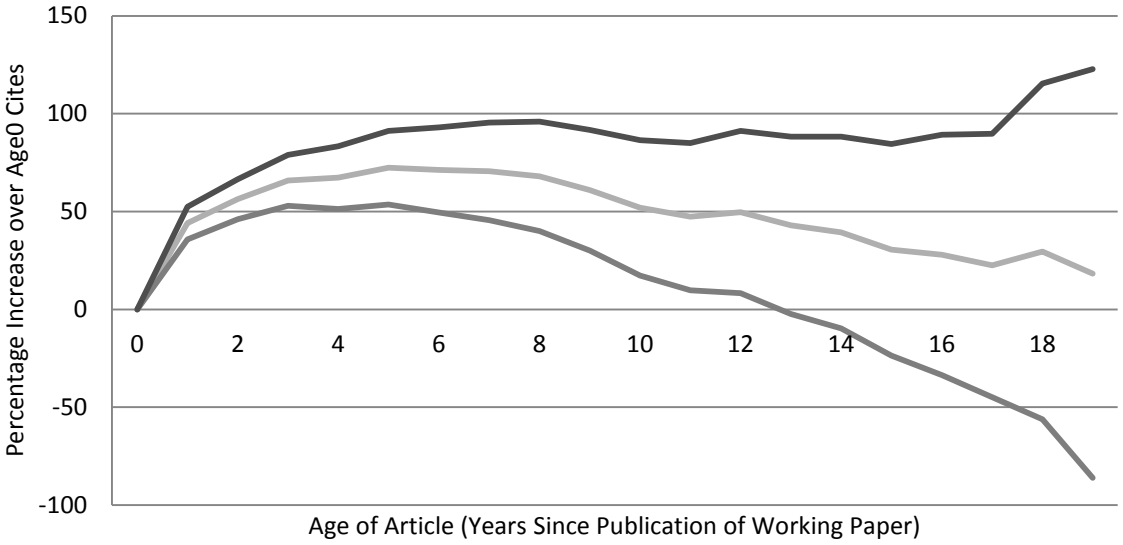
where $cite\_count_{it}$ denotes citations to article i in citing year t, α_i is an article fixed effect, δ_{kt} is a publication dummy (possibly article type k-specific), that equals 1 once an article i is published, x_{it} is a vector of regressors and β is a vector of parameters to be estimated.

Figures 1 and 2 illustrate patterns in citations, which, while interesting in their own right, are important to account for in our estimation procedure. Figure 1 plots the profile of citations over the lifespan of the average article in our sample. Citations peak in the fifth year after first being posted in RePEc. After that, citations gradually fall each year. The 95% confidence interval shows that the estimates are precise early on in the life cycle but become noisier with age. Figure 2 plots secular trends in citations. Citations follow a steady upward trend through 2009, and then plateau through 2015, reaching a level more than 100% higher than in the base year of 1996.

policy, the aggregate effects are essentially unchanged. However, this restricted dataset is too sparse to estimate the more flexible model involving article types.

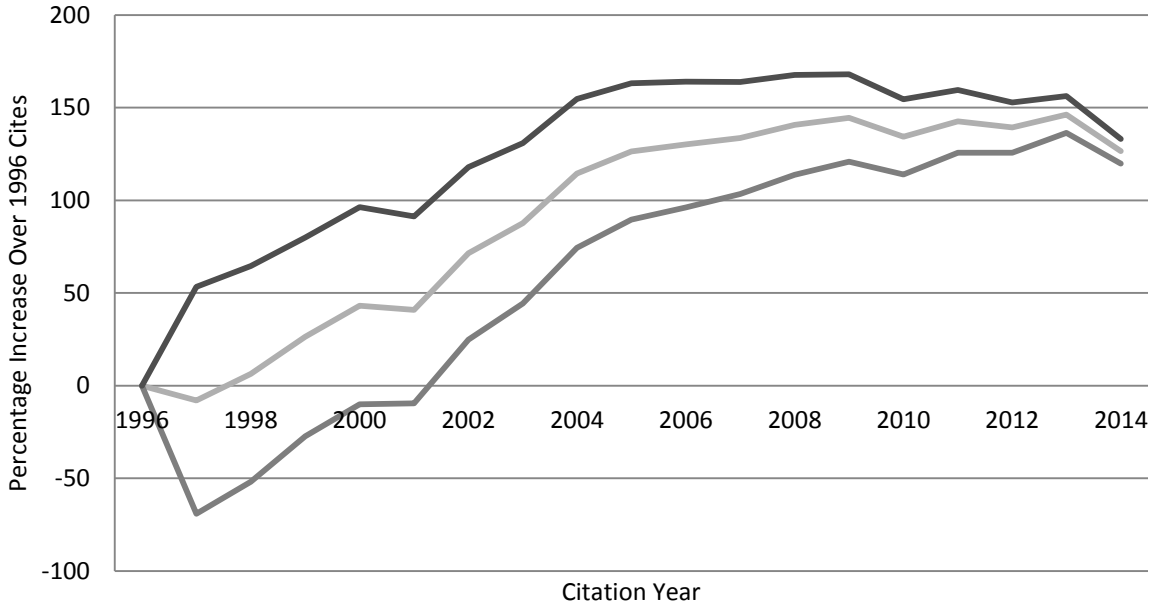
[14] We use Simcoe (2008)'s implementation of this estimator in STATA.

FIGURE 1. CITATION AGE PROFILE



Notes: The middle curve plots a set of fixed age effects from Wooldridge's (1999) PQML procedure. We use Simcoe (2008)'s *xtpqml* command implemented in STATA. The regression also includes citation-year and article fixed effects and binary variables for journal publication and data disclosure. Outside lines bound the 95% confidence interval based on robust standard errors clustered by article.

FIGURE 2. SECULAR TREND IN CITATIONS



Notes: The middle curve plots a set of fixed citation-year effects from Wooldridge's (1999) PQML procedure. We use Simcoe (2008)'s *xtpqml* command implemented in STATA. The regression also includes age and article fixed effects and binary variables for journal publication and data disclosure. Outside lines bound the 95% confidence interval based on robust standard errors clustered by article.

Aspects of article quality varying with article age are captured by including a flexibly specified age profile,

$$\gamma_{1j}AGE_{ijt} + \gamma_{2j}AGE_{ijt}^2$$

where $AGE_{ijt} = t - wp(i)$ is the age of article *i* published in journal *j* in citation year t (relative to the corresponding working paper's date of publication, $wp(i)$), and γ_{1j} and γ_{2j} are coefficients that are allowed to vary across journals.[15]

To control for secular trends we include fixed effects for publication year x citation year interactions (time effects). This set of citation-publication-year interactions is flexible enough to allow each publication year to have a different secular trend and for each secular trend to have an arbitrary pattern.[16] In the regressions, we cluster robust standard errors at the article level.[17]

The most important regressors in x_{it} are the variables of interest:

1. The publication dummy, *Publication* (equal to 1 once a working paper has been published, starting with the year of publication),
2. The interaction of *Publication* with the two article type dummies (*Purely_Empirical*, and *Hybrid*): *Publication_Purely_Empirical* and *Publication_Hybrid*.
3. The data disclosure treatment dummy, *Data_Disclosure* (equal to 1 if data is posted on a journal's website, starting with a working paper's year of publication), and

[15] Our assumption that a journal's age profile extends back in time to the working paper version of published articles implies that these papers are not rejected by any journals in the interim. This is likely the case for the majority of top 4 journal articles. We also estimate (1) without the age profile terms.

[16] When we allow publication and data disclosure effects to vary by article type, we specify article-type-specific citation-publication-year interactions as well.

[17] Our data consists of 4 journals and 838 articles. Following Cameron and Miller (2015), too few clusters can generate overly precise results. When clustering at the journal level we observed this tendency. Therefore, we report results based on clustering at the article level; the corresponding standard errors are substantially larger.

4. The interaction of *Data_Disclosure* with the two article type dummies (*Purely_Empirical*, and *Hybrid*): *Data_Disclosure_Purely_Empirical* and *Data_Disclosure_Hybrid*.

4.3. Descriptive Statistics

Table 1 provides descriptive statistics of the variables under study at the article-citing year level.

Table 1 | Descriptive statistics (Article citing-year level)

	Mean	Std. Dev.	Min.	Max.	Obs.
Dependent variable					
Citation_Count	13.22	18.540	0	276	9,895
Period under study					
Citation_Year	2009	4.340	1996	2015	9,895
Article types					
Purely_Empirical	0.219	0.413	0	1	9,895
Hybrid	0.781	0.413	0	1	9,895
Main variables of interest (data disclosure treatment)					
Data_Disclosure	0.310	0.463	0	1	9,895
Data_Disclosure_Purely_Empirical	0.067	0.250	0	1	9,895
Data_Disclosure_Hybrid	0.243	0.429	0	1	9,895
Article characteristics					
Publication	0.764	0.425	0	1	9,895
Publication_Purely_Empirical	0.173	0.378	0	1	9,895
Publication_Hybrid	0.592	0.492	0	1	9,895
Publication_Year_Journal	2006	3.704	2000	2012	9,895
Publication_Year_WP	2003	3.779	1996	2010	9,895

Notes: Years and volumes when mandatory data disclosure policy was implemented by journal: AER: 2005, Vol. 95; Econometrica: 2004, Vol. 72; JPE: 2005, Vol. 113; ReStud: 2006, Vol. 73.

Our dependent variable is $Citation\_Count_{it}$ which indicates cites to article i in citing year t, as given by $Citation\_Year_{it}$, with t=1996,…,2015. The articles in our sample receive, on average, 13.2 cites in a given citing year. Hybrid and purely empirical articles represent 78.1% and 21.9% of the observations, respectively. We find that average citation counts for purely empirical and hybrid articles differ. On average, purely empirical articles in our sample receive 13.91 cites in a given citing year while hybrid

articles receive 13.03 cites (results not reported in Table 1). Overall, data disclosure is observed in 31.0% of the sample (*Data_Disclosure_Purely_Empirical*: 6.7%; *Data_Disclosure_Hybrid*: 24.3%).

Table 1 also reports some useful summary article-level statistics. Articles were published in one of four journals between 2000 and 2012. Corresponding working papers were published in RePEc, on average, about three years before the respective journal publication. Working papers appeared between 1996 and 2010. If we divide each article's citing year time series into pre- and post-journal publication periods, then we find that the post-publication citing years account for 76.4% of our observations. The remaining 23.6% are citing years to working papers.

5. Empirical Analysis

5.1. Empirical Articles

Table 2 reports marginal effects for PQML regressions based on the simplest versions of (1). In each specification, we include fixed effects for publication *x* citation year interactions (time effects) and article fixed effects. In specification (1), we separately estimate the aggregate citation impact of publication and data disclosure, respectively. In specification (2), we include journal-specific citation age profiles. In specification (3), we allow for the possibility that both the impact of publication and data disclosure may vary across article types. Specification (3) is the basis for specification (4) where we include journal-specific citation age profiles.

First, our evidence for a positive open data citation effect is weak. In column (1), we provide evidence for a modest general data disclosure citation effect of 12.9% that is statistically significant at the 10% level. This effect decreases in magnitude (6.4%) and precision once we include journal-specific citation age profiles in (2). When we interact *Data_Disclosure* and *Publication* with article types, a similar story emerges. *Data_Disclosure_Hybrid* is positive and significant at the 10% level in (3); the estimate implies a citation impact of 13.7%. The corresponding estimate for *Data_Disclosure_Purely_Empirical* is positive but insignificant. It appears that data disclosure may only benefit hybrid articles. However,

once we include age profiles in (4) the citation effect of *Data_Disclosure_Hybrid* decreases in magnitude to 7.1% and is no longer significant; the open data citation effect for purely empirical articles is small, slightly negative and insignificant.

Table 2 | Citation effect of data disclosure

VARIABLES	(1) PQML	(2) PQML	(3) PQML	(4) PQML
Publication	0.264***	0.310***		
	(0.066)	(0.069)		
Publication_Purely_Empirical			0.386**	0.462***
			(0.196)	(0.201)
Publication_Hybrid			0.249***	0.297***
			(0.073)	(0.079)
Data_Disclosure	0.129*	0.064		
	(0.074)	(0.070)		
Data_Disclosure_Purely_Empirical			0.067	-0.023
			(0.152)	(0.136)
Data_Disclosure_Hybrid			0.137*	0.071
			(0.084)	(0.080)
Observations	9,895	9,895	9,895	9,895
Number of article groups	838	838	838	838
Article fixed effects	Yes	Yes	Yes	Yes
Time effects [a]	Yes	Yes	Yes	Yes
Quadratic age profile	No	Yes	No	Yes
Wald Test Statistics, Vars.[b]				
Chi-squared			0.46	0.63
Degrees of freedom			1	1
p-value			0.497	0.428
Wald Test Statistics, Vars.[c]				
Chi-squared			0.15	0.35
Degrees of freedom			1	1
p-value			0.702	0.557

Notes: Cites to article *i* in citing year *t* dependent variable. Results from Wooldridge's (1999) PQML procedure. We use Simcoe (2008)'s *xtpqml* command implemented in STATA. Results converted into marginal effects. Marginal effects are given by exp(*b*)-1, where *b* is the regression coefficient and exp(*b*) is the incidence rate ratio. Robust standard errors (clustered at the article level) reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

[a] One set of common publication year *x* citation year interaction terms included in (1) and (2). Two sets of interaction terms included in (3) and (4), corresponding to each article type.

[b] Variables: *Publication_Purely_Empirical* and *Publication_Hybrid.*

[c] Variables: *Data_Disclosure_Purely_Empirical* and *Data_Disclosure_Hybrid.*

Second, publication in a journal has a positive and statistically significant aggregate citation effect in specifications (1) and (2), ranging from 26.4 to 31%. A more nuanced picture emerges once we allow that publication effects may vary across article types. The citation effect of publication is 55% larger for purely empirical articles compared to hybrid articles, i.e. 38.6% vs. 24.9% in (3) and 46.2% vs. 29.7% in (4); note that these differences are not significant at the 5% level. However, five of the six individual publication citation effects are significant at the 1% level, with the sixth at the 5% level.

5.2. Sensitivity Analysis

So far, our analysis has relied on the use of empirical articles. It is reasonable to assume that empirical articles are the best controls for other empirical articles. Furthermore, accounting for article type within the set of empirical articles reveals additional insights, e.g. the differences observed in the citation impact of publication. Here we add purely theoretical articles for 2 reasons. First, we are interested in estimating the citation benefit of publication for these articles. Second, these articles can act as additional, albeit imperfect controls. Appendix B provides descriptive statistics at the article citing-year level including the observations for purely theoretical articles. The purely theoretical articles now constitute 45.4% of the observations, and empirical articles represent a 54.6% share. Notably, the articles in this augmented sample receive, on average, about 2.2 cites less than those articles in the sample excluding pure theory articles (10.99 vs. 13.2, see also Table 1).

Regression results based on this augmented dataset are reported in Table 3. Since citations to pure theory articles should not benefit directly from the introduction of journal data-disclosure policies, the primary direct effect of adding these additional articles will depend on the relative benefits of publication. For example, if pure theory articles benefit *less* from publication than their empirical counterparts do, then the *aggregate* citation impact of publication (reported in columns 1 and 2) should *decrease* in size. Since publication and data disclosure occur simultaneously, a decrease in the magnitude of the publication parameter is likely to result in an *increase* in the data disclosure terms (in

columns 1 and 2). That is, given this pair of parameters, adding the theoretical articles forces the regression procedure to shift some empirical article "weight" from one parameter to the other. The direction of this shift depends on the relative magnitude of the publication citation effect associated with theoretical articles.

However, once we allow the publication and data disclosure effects (as well as the citation-publication-interaction terms) to vary by article type (columns 3 and 4), the addition of the theoretical articles should have minimal impact on the purely empirical and hybrid article-specific parameters.

The point estimates reported in Table 3 (columns 3 and 4) indicate that publication citation benefits are increasing in the importance of empirical content. For example, using the values in column 4, and relative to the citation impact of a theory article, the advantages enjoyed by hybrid and purely empirical articles are about 50% and 100% larger, respectively. Their relative size helps to explain why the publication dummies in columns 1 and 2 are some 30% smaller than their counterparts in Table 2. Although these differences are not significant at the 5% level, each of the individual publication dummies are statistically significant at the 5% level or better.

Regarding the impact of data disclosure: the point estimates of the *aggregate* data disclosure parameters in Table 3 (columns 1 and 2) are much larger in magnitude than the corresponding parameters in Table 2 (columns 1 and 2), and more precisely estimated. This is expected, given the relatively low citation impact of publishing a theory article. However, once we allow the parameters to vary by article type, the data disclosure parameters in columns 3 and 4 of Table 3 are essentially unchanged compared to the corresponding estimates in Table 2.

Table 3 | Citation effect of data disclosure (incl. purely theoretical articles)

VARIABLES	(1) PQML	(2) PQML	(3) PQML	(4) PQML
Publication	0.178***	0.213***		
	(0.048)	(0.048)		
Publication_Purely_Empirical			0.386**	0.439***
			(0.196)	(0.201)
Publication_Hybrid			0.275***	0.321***
			(0.079)	(0.085)
Publication_Purely_Theoretical			0.183***	0.217***
			(0.072)	(0.069)
Data_Disclosure	0.226***	0.173***		
	(0.062)	(0.061)		
Data_Disclosure_Purely_Empirical			0.070	-0.006
			(0.152)	(0.139)
Data_Disclosure_Hybrid			0.139*	0.085
			(0.087)	(0.084)
Observations	17,135	17,135	17,135	17,135
Number of article groups	1,408	1,408	1,408	1,408
Article fixed effects	Yes	Yes	Yes	Yes
Time effects [a]	Yes	Yes	Yes	Yes
Quadratic age profile	No	Yes	No	Yes
Wald Test Statistics, Vars.[b]				
Chi-squared			1.43	1.71
Degrees of freedom			2	2
p-value			0.488	0.426
Wald Test Statistics, Vars.[c]				
Chi-squared			0.15	0.30
Degrees of freedom			1	1
p-value			0.700	0.581

Notes: Cites to article *i* in citing year *t* dependent variable. Results from Wooldridge's (1999) PQML procedure. We use Simcoe (2008)'s *xtpqml* command implemented in STATA. Results converted into marginal effects. Marginal effects are given by exp(*b*)-1, where *b* is the regression coefficient and exp(*b*) is the incidence rate ratio. Robust standard errors (clustered at the article level) reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

[a] One set of common publication year *x* citation year interaction terms included in (1) and (2). Three sets of interaction terms included in (3) and (4), corresponding to each article type.

[b] Variables: *Publication_Purely_Empirical, Publication_Hybrid* and *Publication_Purely_Theoretical.*

[c] Variables: *Data_Disclosure_Purely_Empirical* and *Data_Disclosure_Hybrid.*

6. Discussion and Policy Implications

There are various stakeholders in the debate surrounding open data: policy-makers, research funders, universities, journal editors and academic authors. The incentives of these stakeholders are not well aligned. On the one hand, policy-makers, research funders, institutions and journals, for a variety of complementary reasons outlined in the introduction, have an interest in supporting open data. Authors, on the other hand, have conflicting incentives, and this is reflected in their behavior: they typically do not share their data (Andreoli-Versbach and Mueller-Langer, 2014). Mandatory data disclosure implies forcing them to do something they would normally not do. Our evidence of negligible aggregate open data benefits may provide an additional explanation for why empirical researchers are typically reluctant to disclose their data voluntarily. Finally, the benefits of open data may also depend on the actual quality of the data and code used in empirical work.

Against this background, we first briefly discuss the implications of our findings for journals and authors. Then, we discuss the use of downloads of data and code as a more direct measure of interest in a paper's data than citations.

6.1. Implications for Journals

In our dataset, citations to empirical articles exceed those to theoretical articles (during the period 1996-2015). Our econometric analysis indicates that citation effects due to publication vary by article type; these publication effects are greater for empirical articles, particularly for purely empirical articles. These results are consistent with Angrist et. al.'s (2017) findings that (1) the empirical share of citations from top economics journals has increased by about 20 percentage points (during the period 1980-2015), and (2) empirical papers now receive more citations than theoretical papers published in the same journal. Angrist and Pischke (2010) offer one possible explanation for these general trends. Namely, the growing importance of empirical work reflects an increase in quality of the underlying research, e.g. the broader use of randomized trials and quasi-experimental methods. However, among empirical papers why do purely empirical papers appear to enjoy a post-publication citation advantage? One

possibility is that less technical papers are more accessible to a wider audience within and outside of economics.

6.2. Implications for Authors

The weak effect of data disclosure may be another reason why empirical researchers are reluctant to voluntarily share their research data (in addition to those mentioned in the introduction, namely the higher cost of data creation, the higher risk of negative replication, and the loss of competitive advantage). In the light of these arguments, mandatory data disclosure may have unintended consequences. For instance, it may cause authors to "flee" to journals that do not require data disclosure.[18]

6.3. Quality of Data and Code

The benefits of open data likely depend on the actual quality of the data and code used in empirical work. Ideally, we would account for these factors in our regressions. We contacted the journals under study in several waves of emails from December 2017 to February 2018 and requested data on data and code downloads for the articles in our sample. Downloads are probably a more direct measure of interest in a paper's data than citations. Indeed, the number of citations to an empirical paper is an imperfect measure of interest in the paper's data.

We did not obtain the requested data for the following reasons. Three of the four journals under study could not provide us with data on data downloads either because the data did not exist at the time of our request, i.e., the journals did not track the usage of data files, or because it would have been too costly for the journal to provide it. Notably, one journal (ReStud) did provide us with the download metrics.

[18] Note that – with the exception reported in footnote 11 – the top economics journals were the first to adopt mandatory data disclosure policies (Econometrica: 2004; AER: 2005; JPE: 2005; ReStud: 2006). Arguably, the probability that authors might "flee" to other less demanding journals is lowest for the top journals. However, it would be interesting to examine whether QJE received more submissions of empirical papers after the other four top-5 journals introduced their data policies.

However, ReStud only began to track usage of data files in January 2017, well outside of our 1996-2015 sample period.

So exploring these possible effects is a task for future research once the necessary data becomes available. The fact that data editors have been recently appointed by leading economics associations for their respective journals, including the *American Economic Review* and the *Canadian Journal of Economics*, is a promising development (Duflo and Hoynes, 2018; Canadian Economics Association, 2016).

7. Conclusion

In this paper, we examine the citation effect of (mandatory) data disclosure on empirical articles published in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, and the *Review of Economic Studies*. We exploit a natural experiment in which data is disclosed by slightly more than half (54%) of the empirical articles in our sample. We use data provided by RePEc that tracks citations to papers before and after publication in a journal. This approach allows us to identify the separate citation effects of publication and data disclosure while controlling for the intrinsic quality of each article.

Evidence for a positive open data citation effect is weak. The aggregate citation impact of mandatory data disclosure among empirical articles is 6% but is not statistically significant. On the other hand, the citation impacts of publication are substantial and precisely estimated. Pure theory and hybrid articles enjoy citations benefits of 22% and 32%, respectively. The benefits for purely empirical articles are 44%.

From a science policy perspective, the incentives to support open data of the different stakeholders are not necessarily well aligned. While policy-makers, research funders, institutions and journals have an interest in supporting open data, mandatory data disclosure implies forcing authors to do something they would normally not do. Our evidence of negligible aggregate open data benefits may provide an

additional explanation for why empirical researchers are typically reluctant to disclose their data voluntarily.

Finally, there are potential benefits of open data that are beyond the scope of the present paper. For instance, data disclosure could increase the credibility and/or understanding of empirical economic research for scholars outside of economics. That is, data disclosure may contribute to the "credibility revolution" in empirical economic research (Angrist and Pischke, 2010; Hamermesh, 2013 & 2018). Data disclosure makes replication more feasible because it reduces its cost (Kiri et al., 2018; Lacetera and Zirulia, 2011). In future work, it would be interesting to examine the effect of replication on citations to replicated articles.[19]

[19] In our sample, only 15 of the 838 empirical articles under study were eventually replicated (1.6%). We do not have sufficient variation in terms of replication in our sample to conduct a meaningful analysis. Typically, replication is very rare in economics (Mueller-Langer et al., 2019). Only 0.1% of articles published in the top-50 economics journals from 1974 to 2014 were eventually replicated in a published replication study.

References

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: An empirical investigation. American Economic Review, 91(5), 1369-1401.

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., Ioannidis, J.P.A., 2011. Public availability of published research data in high-impact journals. PLoS ONE, 6, e24357.

Altman, M., King, G., 2007. A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13 (3/4), Retrieved from http://www.dlib.org/dlib/march07/altman/03altman.html (accessed: 29 January 2019).

Allan, R., 2012. Editorial: Geoscience data. Geoscience Data Journal, Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/gdj.3/full (accessed: 10 November 2017).

American Economic Review, 2019. AER Data Availability Policy, Retrieved from http://www.aeaweb.org/aer/data.php (accessed: 29 January 2019).

Anderson, R.G., Greene, W.H., McCullough, B.D., Vinod, H.D., 2008. The role of data/code archives in the future of economic research. Journal of Economic Methodology, 15(1), 99-119.

Andreoli-Versbach, P., Mueller-Langer F., 2014. Open access to data: An ideal professed but not practised. Research Policy, 43(9), 1621-1633.

Angrist, J.D., Bettinger, E., Bloom, E., King, E., Kremer, M., 2002. Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. American Economic Review, 92(5), 1535-1558.

Angrist, J.D., Azoulay, P., Ellison, G., Hill, R., Lu, S.F., 2017. Inside job or deep impact? Using extramural citations to assess economic scholarship. NBER Working Paper No. 23698.

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

Begley, C.G., Ellis, L.M., 2012. Drug development: Raise standards for preclinical cancer research. Nature, 483, 531-533.

Bell, A., Johnston, R., Jones, K., 2015. Stylised fact or situated messiness? The diverse effects of increasing debt on national economic growth. Journal of Economic Geography, 15(2), 449-472.

Bernanke, B.S., 2004. Editorial statement. American Economic Review, 94 (1), 404–404.

Burgelman, J.-C., Osimo, D., Bogdanowicz, M., 2010. Science 2.0 (change will happen ...), First Monday, 15(7), 5 July 2010.

Camerer, C.F., Dreber, A., Forswell, E. et al., 2016. Evaluating replicability of laboratory experiments in economics. Science, 10.1126/science.aaf0918.

Cameron, A.C., Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50(2), 317-372.

Canadian Economics Association (2016), Annual General Meeting Minutes, 5 June 2016, University of Ottawa, Ottawa, ON.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. Journal of Political Economy, 113(1), 1-45.

Costello, M.J., 2009. Motivating online publication of data. BioScience, 59(5), 418-427.

Dasgupta, P., David, P.A., 1994. Toward a new economics of science. Research Policy, 23(5), 487-521.

Dewald, W.G., Thursby, J.G., Anderson, R.G., 1986. Replication in empirical economics: The Journal of Money, Credit and Banking Project. American Economic Review, 76(4), 587-603.

Doldirina, C., Friis-Christensen, A., Ostlaender, N. et al., 2015. JRC data policy, JRC Technical Report EUR 27163.

Dorch, B., 2012. On the citation advantage of linking to data: Astrophysics. Retrieved from http://hprints.org/hprints-00714715/ (accessed: 29 January 2019).

Duflo, E., Hoynes, H., 2018. Report of the search committee to appoint a data editor for the AEA. AEA Papers and Proceedings, 108, 745.

Economic and Social Research Council, 2019. ESRC Research Data Policy, Retrieved from http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx (accessed: 29 January 2019).

Ellison, G., 2002. The slowdown of the economics publishing process. Journal of Political Economy, 110(5), 947-993.

European Commission, 2012. Towards better access to scientific information: Boosting the benefits of public investments in research. European Commission: Brussels.

European Commission, 2016. H2020 Programme: Guidelines on open access to scientific publications and research data in Horizon 2020, European Commission, Directorate-General for Research & Innovation: Brussels.

European Commission, 2017. EOSC Declaration, European Commission, Directorate-General for Research & Innovation: Brussels.

Feigenbaum, S., Levy, D.M., 1993. The market for (ir)reproducible econometrics. Social Epistemology, 7(3), 215-232.

Furman, J.L., Stern, S., 2011. Climbing atop the shoulders of giants: The impact of institutions on cumulative research. American Economic Review, 101(5), 1933-1963.

Glandon, P., 2011. Appendix to the report of the editor: Compliance project report on the American Economic Review data availability compliance project, American Economic Review Papers and Proceedings, 101(3), 659-699.

Gleditsch N.P., Metelits, C., Strand H., 2003. Posting your data: Will you be scooped or will you be famous? International Studies Perspectives, 4(1), 89-97.

Hamermesh, D.S., 1997. Some thoughts on replications and reviews. Labour Economics, 4(2), 107-109.

Hamermesh, D.S., 2007. Viewpoint: Replication in economics. Canadian Journal of Economics, 40(3), 715–733.

Hamermesh, D.S., 2013. Six decades of top economics publishing: Who and how? Journal of Economic Literature, 51(1), 162-172.

Hamermesh, D.S. 2018. Citations in economics: Measurement, uses and impacts. Journal of Economic Literature, 56(1), 115-156.

Henneken, E.A., Accomazzi, A., 2011. Linking to data - effect on citation rates in astronomy. Retrieved from https://arxiv.org/abs/1111.3618 (accessed: 29 January 2019).

Herndon, T., Ash, M., Pollin, R., 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge Journal of Economics 38, 257-279.

Kiri, B., Lacetera, N., Zirulia, L, 2018. Above a swamp: A theory of high-quality scientific production. Research Policy, 47(5), 827-839.

Kling, J. R., Liebman, J. B., Katz, L. F., 2007. Experimental analysis of neighborhood effects. Econometrica, 75(1), 83-119.

Lacetera, N., Zirulia, L., 2011. The economics of scientific misconduct. Journal of Law, Economics, and Organization, 27(3), 568-603.

Levelt, Noort and Drenth Committees, 2012. Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. Retrieved from https://www.commissielevelt.nl/ (accessed: 29 January 2019).

McCabe, M.J., Snyder, C.M., 2015. Does online availability increase citations? Theory and evidence from a panel of economics and business journals. Review of Economics and Statistics, 97(1), 144-165.

McCabe, M.J., Snyder, C.M., 2014. Replication data for: Does online availability increase citations? Theory and evidence from a panel of economics and business journals. doi:10.7910/DVN/26904, Harvard Dataverse, V2.

McCullough, B.D., 2009. Open access economics journals and the market for reproducible economic research. Economic Analysis and Policy, 39(1), 117-126.

McCullough, B.D., McGeary, K.A., Harrison, T.D., 2008. Do economics journal archives promote replicable research? Canadian Journal of Economics, 41(4), 1406-1420.

McCullough, B.D., McGeary, K.A., Harrison, T.D., 2006. Lessons from the JMCB Archive. Journal of Money, Credit and Banking, 38(4), 1093-1107.

McCullough, B. D., Vinod, H. D., 2003. Verifying the solution from a nonlinear solver: A case study. American Economic Review, 93(3), 873-892.

Mirowski, P., Sklivas, S., 1991. Why econometricians don't replicate (although they do reproduce). Review of Political Economy, 3(2), 146-163.

Mueller-Langer, F., B. Fecher, D. Harhoff, Wagner, G.G., 2019. Replication studies in economics: How many and which papers are chosen for replication, and why? Research Policy, 48(1), 62-83.

National Institutes of Health, 2003. NIH Data Sharing Policy, Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed: 29 January 2019).

National Science Foundation, 2014. Proposal and Award Policies and Procedures Guide, Retrieved from https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_print.pdf (accessed: 29 January 2019).

Nature, 2009. Editorial, Data's shameful neglect. Nature, 461, 145.

Nature, 2013. Editorial announcement: launch of an online data journal. Nature 502(7470), 142.

Nelson, B., 2009. Empty archives. Nature, 461, 160-163.

Organisation for Economic Co-operation and Development, 2007. OECD principles and guidelines for access to research data from public funding. OECD: Paris.

Piwowar, H.A., Day, R.S., Fridsma, D.B., 2007. Sharing detailed research data is associated with increased citation rate. PLoS ONE, 2(3), e308.

Piwowar, H. A., Vision, T. J., 2013. Data reuse and the open data citation advantage. PeerJ, 1, e175.

Reinhart, C. M., V. R. Reinhart, Rogoff, K. S., 2012. Public debt overhangs: Advanced-economy episodes since 1800. Journal of Economic Perspectives, 26(3), 69-86.

Reinhart, C. M., Rogoff, K. S., 2010a. Growth in a time of debt. National Bureau of Economic Research. At http://www.nber.org/papers/w15639.pdf (accessed: 29 January 2019).

Reinhart, C. M., Rogoff, K. S., 2010b. Growth in a time of debt. American Economic Review Papers and Proceedings, 100(2), 573-578.

Reinhart, C.M., Rogoff, K.S., 2013. Errata: Growth in a time of debt. Harvard University. Retrieved from http://scholar.harvard.edu/rogoff/publications/growth-time-debt (accessed: 29 January 2019).

Sears, J. R., 2012. Data sharing effect on article citation rate in paleoceanography. IN53B-1628. AGU Fall Meeting 2011. Retrieved from http://static.coreapps.net/agu2011/html/IN53B-1628.html (accessed: 29 January 2019).

Simcoe, T., 2008. XTPQML: Stata module to estimate fixed-effects Poisson (Quasi-ML) regression with robust standard errors. Statistical Software Components, Boston College Department of Economics. Retrieved from: econpapers.repec.org/RePEc:boc:bocode:s456821 (accessed: 29 January 2019).

Stephan, P.E., 1996. The economics of science. Journal of Economic Literature, 34(3), 1199-1235.

US House of Representatives, 2007. America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Act, 110th Congress, 1st Session. H.R. 2272, Section 1009, Release of scientific research results, Government Printing Office: Washington.

Vlaeminck, S., Wagner, G.G., Wagner, J., Harhoff, D., Siegert, O., 2013. Replizierbare Forschung in den Wirtschaftswissenschaften erhöhen: Eine Herausforderung für wissenschaftliche Infrastrukturdienstleister. LIBREAS. Library Ideas # 23: Forschungsdaten. Metadaten. Noch mehr Daten. Forschungsdatenmanagement, 29-42.

Wooldridge, J.M., 1999. Distribution-free estimation of some nonlinear panel data models. Journal of Econometrics, 90(1), 77-97.

Appendix A Classifying Articles

A.1 Overview

We use an automated classification process to determine whether an article is of purely theoretical nature, purely empirical nature or has elements of both styles (henceforth, hybrid articles). To set-up this process we use a sample of 6,436 articles that are published in economics journals between 1995 and 2015 and for which pre-prints are available on RePEc (henceforth, population). Table A1 (Appendix) provides an overview of the journals where these articles are published. Note that all 1,408 top 5 journal articles that we examine in the paper are included in the population of 6,436 articles. The 6,436 PDFs obtained from RePEc are at the core of our classification process which is based on the frequency of style-specific words as they appear in the articles under study. Using this classification process, we classify 6,436 articles into the three respective styles, i.e., purely theoretical, purely empirical or hybrid.

Generally, the idea is to set up a relationship between a set of style-specific words within an article and its style. The dependency of article style on the set of specific words within an article is estimated by probit regressions. However, in the final outcome of the classification we will focus on two styles only, i.e., purely theoretical article vs. empirical (hybrid and purely empirical) articles as the estimation of the bilateral distinction yields results with a higher accuracy. That is, we will use the automated classification process to distinguish between purely theoretical and empirical articles. However, in order to distinguish between hybrid and purely empirical articles *within* the set of empirical articles we rely on the more precise manual classification procedure.

In addition to the RePEc articles, we use articles published in two theory journals and one empirical journal. We use two samples. The first sample consists of 200 randomly chosen and manually classified articles taken from the population of 6,436 RePEc articles. The second sample is a training sample consisting of 2,619 economics articles (see section A2 “Training data” below). In the regressions, the dependent variable is article style and the independent variables are the word counts of 48 pre-selected words within each article. The idea is that the word composition within each article will characterize the

article regarding its style. From the coefficients, which we receive by regressing the (manually) pre-determined article styles on their corresponding word counts, we predict the styles of the rest of the population.

Our objective is to predict the style of a given article with a rate of correct classification of at least 95%. In order to test this we take a random sample of 100 articles out of the population. This sample is manually classified and not used in the ultimate estimation of coefficients. We assess the rate of correct prediction by comparing the known (manually classified) style to the predicted style of the articles in this test sample. As we will see, the bilateral distinction between purely theoretical and empirical articles performs with an accuracy of 96% for this random sample of 100 articles.

A.2 Training Data

Figure A1 provides an overview of the classification process and the different data samples used in this process. In the following, we describe the creation of the different samples in more detail.

Manually classified estimation sample and test sample

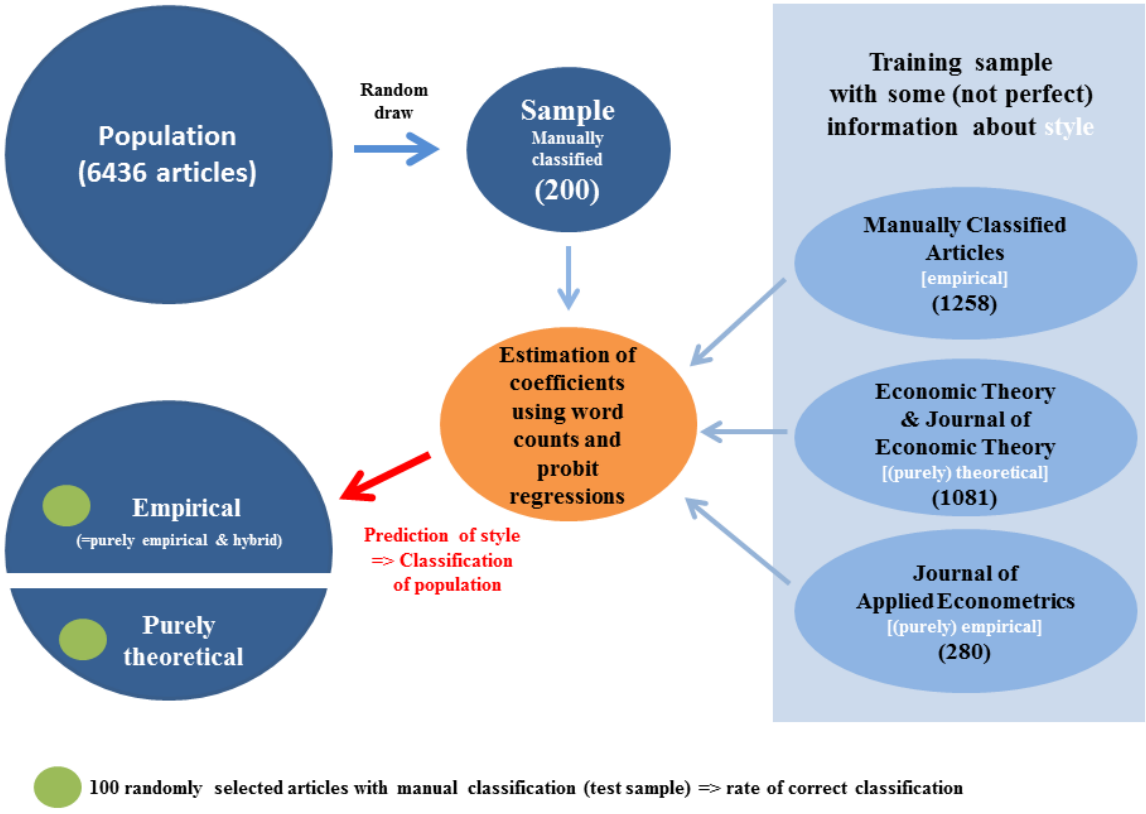
The manual style classification works as follows: We classify an article as purely empirical if the main contribution of the author(s) is derived by the use of data and empirical estimation techniques. We classify an article as purely theoretical if the authors derive their results exclusively via mathematical and axiomatic deductions, i.e., they do not use any data to derive their results. We classify an article as hybrid if it contains elements of both "pure" styles.

We draw a random sample of 200 articles out of the population of 6,436 RePEc articles and manually classify those articles into the three styles (see Figure A1). This sample is the basis for the probit regression used to estimate the article styles for the whole population.

We draw another random sample of 100 articles out of the population and classify them manually. In a final step, this sample will be used to compare the predicted style to the manually classified style. Therefore, we refer to them as test sample (see bottom of Figure A1). Since these 100 articles are used

to check the accuracy of our regression model, their manually classified styles will not be used to predict article styles in the estimation process.

Figure A1 | Overview of Classification Process and Data Samples



Training sample

The manual classification of the 200 articles is precise. However, it results in a rather small sample. To get more information on the relationship between article style and the set of style-specific words within a given article we use another sample where we have some (albeit incomplete) information about article styles. Overall, this training sample consists of 2,619 additional articles from three sources (see right-hand side of Figure A1):

1. **1,258** manually classified, **empirical articles** obtained from Mueller-Langer et al. (2019)[20]
2. **1,081** articles published in **theoretical journals,** i.e, *Economic Theory* and *Journal of Economic Theory*[21]
3. **280** articles published in an **empirical journal,** i.e., *The Journal of Applied Econometrics*[22]

For each of these three sources, we can make reasonable assumptions on the respective article styles. Given the journals' focus, we assume that neither *Economic Theory* nor *Journal of Economic Theory* publish purely empirical articles. In addition, we assume that the *Journal of Applied Econometrics* does not publish purely theoretical articles. However, there could be hybrid articles. We make the following assumptions about the existence of hybrid articles in those journals to address this issue:

Restrictive assumption Articles published in *Economic Theory* and *Journal of Economic Theory* are exclusively purely theoretical. Articles published in *Journal of Applied Econometrics* are exclusively purely empirical.

Weak assumption Articles published in *Economic Theory* and *Journal of Economic Theory* are either purely theoretical or hybrid. Articles published in the *Journal of Applied Econometrics* are either purely empirical or hybrid.

In the following we run every step of the estimation process separately for the restrictive assumption and the weak assumption. Thus, after the first run of probit regressions, we will have two separate results, one for each assumption. We then test the performance of the two assumptions in terms of their rate of correct style prediction.

[20] Mueller-Langer et al. (2019) explore articles published in the top-50 economics journal according to Web of Science journal impact factors and manually classify the articles under study. However, Mueller-Langer et al. (2019) do not distinguish empirical articles between purely empirical and hybrid articles.

[21] We retrieve the PDFs of the most recent articles published in the two journals.

[22] We retrieve the PDFs of the most recent articles published in this journal.

A.3 Text Processing

The initial selection of style-specific words is derived from the manual classification of the first sample of 200 articles. From this manual classification, we obtain two sets of theory-style specific and empirical-style specific words (in total, 79 words). For instance, words such as "qed", "proof" or "theorem" do not appear in articles manually classified as purely empirical articles. In contrast, words such as "p-value", "standard error" or "estimate" do not appear in articles manually classified as purely theoretical. These 79 words are automatically counted by a PDF-reading computer software for each of the 9,055 articles of the population and the training sample. During the probit regressions the selection of style-specific words is constantly adjusted. From the full set of 79 words we eliminate words which cause problems due to multicollinearity or have virtually no explanatory power (as given by high *p*-values in the probit regressions). This process results in a final list of the following 48 (stem) words:

*agent, assume, calibrat, condition, control, correlated, data, dependent variable, effect, equation, equilibri, estimate, eviden, examin, experiment, finding, function, index, lemma, model, monthly, numerical, optimal, order, panel, percentage, predict, proof, proposition, p-value, qed, quarterly, regression, remark, representative, sample, significan, solution, standard error, **, stationarity, statistic, table, test, theor, theorem, utility and welfare.*

A.4 Estimation

In the probit regressions, the dependent variable is article style and the independent variables are the word counts of the 48 pre-selected words within each article.

Binary dependent variables

We estimate article style based on how often theory-specific or empirical-work-specific words are used in a given article. For a better econometric performance we split this classification problem into two separate problems. First, we estimate whether an article is purely empirical or hybrid/purely theoretical. Then, we estimate whether an article is purely theoretical or hybrid/purely empirical. Thus, we create binary dependent variables of the style we have so far and conduct two binary choice regressions. First,

we determine the probability of an article to be purely empirical (vs. hybrid/purely theoretical). Second, we determine the probability of an article to be purely theoretical (vs. hybrid/purely empirical).

Based on this, we use two consecutive probit regressions. We create four different binary dependent variables due to the following reasons. We make two assumptions (weak vs. restrictive) about the journal articles and have two probit regressions for estimating and predicting article style.

First round of probit regressions

Initially, we have information about the dependent variable (article style) of the manually classified articles and the articles in the training sample. Using this information, we estimate the coefficients of the independent variables (word counts) and estimate article style for the whole population. In other words, we employ the probit regressions in the first round to define an article to be purely empirical, or purely theoretical, respectively.

We introduce an iterative process at this stage. Articles that surpass a very high threshold in their probability to have a given style are assumed to have that style. In order for us to assume that an article has a particular style, the article must be predicted to have that style with a probability of at least 99% under both the weak and the restrictive assumption. Then information about that article is again used to estimate the coefficients of the independent variables and the style of the rest of the population. This step is repeated 12 times. Each run increases the share of articles with known style and thereby increases the precision of the probit regression. Additionally, in each run the probability threshold will be lowered by 0.5% since we can except to have a more precise estimation.[23]

[23] The number of runs as well as the level and descending order of the threshold has been established by numerous runs of the whole process while checking what performs best in terms of the rate of correct style prediction.

Second round of probit regressions

The second round of Probit regressions assigns a predicted style to *each* article. Obtaining the predicted style for each article, we firstly check the accuracy of the process. Then, we conduct the final classification of the full population.

The restrictive assumption is dropped at this point as the weak assumption performs better in terms of rate of correct style prediction.[24] Therefore, we continue with two dependent variables running two consecutive Probit regressions. The further proceedings are straightforward. We run the two probit regressions, purely empirical vs. hybrid/purely theoretical and purely theoretical vs. hybrid/purely empirical. With the predicted coefficients for each of the words we predict the style of an article according to the following thresholds of predicted probabilities:

- An article with predicted probability above 40% to be purely theoretical and below 40% to be purely empirical is classified as *purely theoretical*.
- An article with predicted probability below 40% to be purely theoretical and above 40% to be purely empirical is classified as *purely empirical*.
- An article meeting none or both criteria to be classified as *purely theoretical* or *purely empirical* is classified as *hybrid*.

The threshold of 40% is obtained from numerous runs while choosing the most constructive threshold in achieving a sufficiently high rate of correct style prediction for the articles in the test sample.

A.5 Classification

The accuracy of the process is measured by comparing the predicted styles and the manually classified styles of the 100 articles of the test-sample. When we consider three styles, i.e., purely theoretical, purely empirical or hybrid, we achieve a rate of correct prediction of 86%. When we consider two

[24] However, the restrictive assumption was needed in the first round to provide clear-cut information on the predicted style as it was a second confirmation of an article being evidently empirical or theoretical.

styles, i.e., purely theoretical and empirical, we achieve a rate of correct prediction of 96%. In the latter case, we obtain the final classification of 1,341 purely theoretical articles and 5,095 empirical articles. Finally, as for the sample of 1,408 articles used in this paper, we obtain a final classification of 570 purely theoretical articles and 838 empirical articles.

Table A1 | Overview of 6,436 economics articles in the population

Journal	Count	Percent
American Economic Journal: Applied Economics	110	1,7%
American Economic Journal: Economic Policy	109	1,7%
American Economic Journal: Macroeconomics	116	1,8%
American Economic Journal: Microeconomics	59	0,9%
American Economic Review	970	15,1%
American Journal of Agricultural Economics	80	1,2%
Applied Economic Perspectives and Policy	6	0,1%
Australian Economic Review	9	0,1%
Brookings Papers on Economic Activity	29	0,5%
Canadian Journal of Economics	157	2,4%
De Economist	17	0,3%
Econometrica	468	7,3%
Economic Journal	505	7,8%
Economics - The Open-Access, Open-Assessment E-Journal	36	0,6%
Empirical Economics	186	2,9%
IMF Economic Review	26	0,4%
IMF Staff Papers	81	1,3%
International Journal of Forecasting	97	1,5%
Journal of Agricultural and Resource Economics	14	0,2%
Journal of Applied Econometrics	257	4,0%
Journal of Business & Economic Statistics	190	3,0%
Journal of Human Resources	160	2,5%
Journal of Labor Economics	193	3,0%
Journal of Law, Economics and Organization	48	0,7%
Journal of Money, Credit and Banking	389	6,0%
Journal of Political Economy	323	5,0%
Journal of the European Economic Association	300	4,7%
Land Economics	29	0,5%
Federal Reserve Bank of St. Louis Review	41	0,6%
Review of Economic Dynamics	264	4,1%
Review of Economic Studies	405	6,3%
Revista de Economia Aplicada	4	0,1%
Southern Economic Journal	56	0,9%
Studies in Nonlinear Dynamics & Econometrics	36	0,6%
The Economic Record	61	0,9%
The Review of Economics and Statistics	533	8,3%
World Bank Economic Review	72	1,1%
Total:	6,436	

Notes: To set-up the automated classification process adopted in this paper, we use a sample of 6,436 articles that are published between 1995 and 2015 and for which pre-prints are available on RePEc. All 1,408 top 5 economics journal articles that we examine in the paper are included in the population of 6,436 articles. The availability of pre-prints is important for our analysis for two main reasons. First, citations to pre-prints provide us with a proxy for the intrinsic quality of articles. Second, the PDFs obtained from RePEc are at the core of the automated classification process based on the frequency of style-specific words as they appear in the articles under study.

Appendix B Descriptive statistics (Article citing-year level, including purely theoretical articles)

	Mean	Std. Dev.	Min	Max	Obs
Dependent variable					
Citation_Count	10.99	18.70	0	429	17,135
Period under study					
Citation_Year	2009	4.391	1996	2015	17,135
Article types					
Purely_Theoretical	0.454	0.498	0	1	17,135
Purely_Empirical	0.126	0.332	0	1	17,135
Hybrid	0.420	0.494	0	1	17,135
Empirical (Purely_Empirical & Hybrid)	0.546	0.498	0	1	17,135
Main variables of interest (data disclosure treatment)					
Data_Disclosure	0.176	0.380	0	1	17,135
Data_Disclosure_Purely_Empirical	0.039	0.193	0	1	17,135
Data_Disclosure_Hybrid	0.137	0.344	0	1	17,135
Article characteristics					
Publication	0.772	0.420	0	1	17,135
Publication_Purely_Empirical	0.100	0.300	0	1	17,135
Publication_Hybrid	0.317	0.465	0	1	17,135
Publication_Purely_Theoretical	0.355	0.478	0	1	17,135
Publication_Year_Journal	2006	3.616	2000	2012	17,135
Publication_Year_WP	2003	3.687	1996	2010	17,135

Notes: Years and volumes when mandatory data disclosure policy was implemented by journal: AER: 2005, Vol. 95; Econometrica: 2004, Vol. 72; JPE: 2005, Vol. 113; ReStud: 2006, Vol. 73.

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub