RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

215

Open Access to Data: An Ideal Professed but not Practised

Patrick Andreoli-Versbach
Frank Mueller-Langer

February 2013

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Open Access to Data: An Ideal Professed but not Practised

Patrick Andreoli-Versbach*, Frank Mueller-Langer**

*International Max Planck Research School for Competition and Innovation (IMPRS-CI) and Ludwig Maximilian University of Munich
**Max Planck Institute for Intellectual Property and Competition Law and IMPRS-CI

February 21, 2013

Abstract

We provide evidence for the status quo in economics with respect to data sharing using a unique data set with 488 hand-collected observations randomly taken from researchers' academic webpages. Out of the sample, 435 researchers (89.14%) neither have a data&code section nor indicate whether and where their data is available. We find that 8.81% of researchers share some of their data whereas only 2.05% fully share. We run an ordered probit regression to relate the decision of researchers to share to their observable characteristics. We find that three predictors are positive and significant across specifications: being full professor, working at a higher-ranked institution and personal attitudes towards sharing as indicated by sharing other material such as lecture slides.

Keywords: Data sharing, data availability, open science

JEL classification: B40, C80, L59

1 Introduction

Hamermesh (2006, p. 715) suggests that "economists treat replication ... as an ideal to be professed but not to be practised". We study the availability of data that permits other researchers and the general public to identify errors, reject or refine theories and reuse the data for subsequent research. This openness to scrutiny and challenge fosters the ability of self-correction in science.[1] Most researchers seem to embrace the idea of self-correction and

[1]The benefits of releasing data are not confined to replication of results or new research. They may also have strong legal implications as, for example, helping to reduce corruption (Ferraz and Finan, 2008) or detect collusion (Christie and Schultz, 1994).

replicability in science associated with data sharing. In practice, however, those advantages often fail to outweigh researchers' costs to create data and the "'supply of' replicable results in economics has been minimal" according to Anderson et al. (2008, p. 99). The aims of this paper are twofold. First, we use a unique hand-collected data set of 488 empirical economists from the top 100 economics departments and top 50 business schools to provide evidence for the status quo of data sharing in economics. Second, we use an ordered probit model to relate observable characteristics of researchers to their decision of whether or not to share their data.

While most of the top-ranked economics journals have recently introduced a data availability policy,[2] the vast majority of journals either do not have a policy that requires authors to share their data or are reluctant to enforce it (McCullough, 2009; McCullough and Vinod, 2003). Also, Anderson et al. (2008) suggest that authors generally hesitate to share their data and code despite their pre-publication commitment to provide this information. This may suggest that editors, referees and readers are confident that the empirical results presented in the papers are always credible and robust. Unfortunately, this is not always the case. Dewald et al. (1986) tried to replicate 54 papers published in the Journal of Money, Credit, and Banking and could only replicate two. Later, McCullough et al. (2006) tried to replicate 69 articles with archived data entries published in the same journal and could only replicate 14. Also, McCullough et al. (2008) tried to replicate 117 articles with archived entries published in the Federal Reserve Bank of St. Louis Review and could only replicate 9. These findings raise concerns regarding the credibility and reliability of empirical work.[3]

In this paper we provide evidence that data sharing in economics is still scarce and that, even though the ease of sharing has increased[4] and many top-ranked journals require sharing,

[2]For instance, the American Economic Journals, Review of Economic Studies, Journal of Political Economy, and Econometrica.

[3]This phenomenon is not limited to economics. In medicine, for example, replication of published preclinical studies often fails (Begley and Ellis, 2012) and most authors fail to fully adhere to data availability policies of high-impact journals (Alsheikh-Ali et al., 2011).

[4]Data repositories as, for instance, the US$200,000 repository at the University of Rochester make it easy for researchers to share their data. However, they lie mostly empty (Nelson, 2009).

researchers publish neither data nor codes on their webpages. We further investigate whether there is a link between observable researchers' characteristics and their decision to share. We believe that our analysis is a first step to understand the incentives to share data and the relation between data sharing and observable characteristics of researchers.

The remainder of the paper is organized as follows. Section 2 describes the data. In Section 3, we run the regressions and present the results. Section 4 concludes our study.

2 Data

In this section we evaluate the hand-collected data on the sharing behaviour of empirical researchers. In addition to an overview of the status quo of data sharing in economics, we relate the researchers' observable characteristics to their decision of whether or not to share their data. In the regression model, we distinguish between structural characteristics, e.g. status, research experience or gender, and individual preferences towards sharing as proxied by sharing material other than data, e.g. lecture slides.

The data set we use consists of 488 observations that were collected from researchers' webpages in the time period from March 2012 to July 2012. In the data set we gathered information on the willingness to share data of 388 economists affiliated with an economics department and 100 affiliated with a business school. The researchers were chosen uniformly across the top 100 economics departments (four observations each) and top 50 business schools (two observations each) and randomly within the respective institution.[5] Economics departments were chosen using the Shanghai Ranking 2011 in Economics and Business[6] and business schools were chosen using the Financial Times Global MBA Ranking 2011.[7] We also collected information on the position, research grants, gender, years since Ph.D. and age of researchers from their CVs, and whether researchers share other material from their

[5] Three business schools (London Business School, INSEAD and HEC Paris) were listed in both rankings. Thus, we excluded 12 observations from the top 100 economics departments in our data set.

[6] See http://www.shanghairanking.com/SubjectEcoBus2011.html (last accessed February 21, 2013).

[7] See http://rankings.ft.com/businessschoolrankings/global-mba-rankings-2011 (last accessed February 21, 2013).

academic webpages. We only collected information on researchers doing mostly empirical work. The sample was constructed to be representative for authors who frequently publish and not for the entire community of researchers. Thus, highly-ranked institutions and more productive researchers are considered more frequently. The reason for this choice is that only authors who publish might share their data. In addition, the work by these authors is arguably more innovative and has had a larger impact on further research (Card and DellaVigna, 2013). For instance, top economics departments hold a dominant position in the production of economics literature (Coupé, 2003). These factors increase the importance of data availability at the top rather than at the bottom of academic research.

Table 1 summarizes the data and shows the correlation coefficients with respect to *ShareData*, the dependent variable in our model. *ShareData* is an ordinal variable that represents the extent to which a researcher makes her data available. It can take three values, no sharing (0), some sharing (1), and full sharing (2). The values of sharing are defined as follows: $ShareData = 1$ indicates that a researcher shares her data in an only fragmentary manner. For most papers neither the data nor the code is available. $ShareData = 2$ indicates that the researcher has a well-designed data and code section and shares or provides additional information on the availability of *most* of the data. While this definition might seem to contrast with "full sharing" note that many data sets cannot be made available.[8] *Shareothermat* is a binary variable that indicates whether a researcher shares other material such as lecture slides. In contrast to the other variables, *Shareothermat* reflects a preference rather than a characteristic of the researcher and thus we run the regressions with and without it. *Ranking* is given by the Shanghai Ranking (Financial Times Global MBA Ranking) for economics departments (business schools). The Shanghai Ranking for economics departments between position 51 and 75 (76 and 100) is aggregated. Thus, we take the mean ranking values 63 and 88, respectively. $FullProfessor$ is a binary variable indi-

[8] A common reason is confidentiality agreements with the firm or institution providing the data. Note that this does not per se make replication impossible. Some research institutes allow replication using their computers and in addition there are free to use cloud-data services. For example runmycode.org guarantees costless replication and protects confidential data.

cating whether or not the researcher is a full professor. *BusinessSchool* is a binary variable indicating whether the researcher works in a business school or in an economics department. *Experience* is defined as the years since Ph.D. If the date of birth was not available we estimated *Age* as 2012-"Year of Ph.D."+30. *NrResearchGrants* indicates the total number of research grants won by the researcher. Finally, *Male*, *USAempl* and *EUempl* are binary variables and equal to 1 if the researcher is male, works in the U.S. or in Europe, respectively.

Table 1: Summary Statistics

	Mean	Median	St. Dev.	Obs.	Min	Max	Share Data-Correlation
Share Data	0.129	0	0.392	488	0	2	1
Share other mat.	0.184	0	0.388	488	0	1	0.235***
Ranking	44.97	40.5	28.15	488	1	88	-0.081*
Full Professor	0.529	1	0.500	488	0	1	0.133***
Business School	0.205	0	0.404	488	0	1	-0.011
Experience	17.95	16	12.28	488	0	50	0.039
Age	47.72	46	12.11	488	29	80	0.038
Nr. Research Grants	6.717	5	8.060	488	0	47	0.001
Male	0.801	1	0.399	488	0	1	0.124***
USA empl.	0.678	1	0.468	488	0	1	0.07
EU empl.	0.168	0	0.374	488	0	1	-0.022
Other empl.	.153	0	.361	488	0	1	-0.067

3 Empirical Results

The status quo in economics is to not share data: 89.14% of researchers do not share their data. About 8.81% (2.05%) of researchers partly (fully) share their data. A plausible reason for this status quo in data sharing has been brought forward by Moffitt (2007), chief editor of the American Economic Review from 2004 to 2010. He states that:

> "Economists call the 'patent' problem the problem that those who put the effort into constructing a data set and writing programs (months of work) have the right to use it for further research for X years."

Intuitively, researchers have little or no incentive to share and are reluctant to create self-induced competition for their own further research.

We use an ordered probit model to estimate the relation between sharing and the set of observable characteristics of researchers presented above. The results are shown in **Table 2.**

Table 2: Regression Analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable:	Share Data: Ordinal=0 no sharing; 1 some sharing; 2 full sharing					
Regression Model	Ordered probit					
Share other mat.	0.799***	0.801***	0.777***			
	(0.171)	(0.171)	(0.173)			
Ranking	-0.00660**	-0.00662**	-0.00532*	-0.00614**	-0.00622**	-0.00481
	(0.00300)	(0.00304)	(0.00318)	(0.00289)	(0.00292)	(0.00307)
Full Professor	0.437***	0.607***	0.591***	0.441***	0.603***	0.602***
	(0.163)	(0.212)	(0.226)	(0.159)	(0.206)	(0.220)
Business School	-0.120	-0.138	-0.0910	-0.195	-0.215	-0.148
	(0.208)	(0.209)	(0.219)	(0.200)	(0.201)	(0.211)
Experience		-0.0101	-0.0133		-0.00968	-0.0137
		(0.00883)	(0.00911)		(0.00851)	(0.00880)
Nr. Research Grants		-0.00593	-0.00681		-0.00568	-0.00649
		(0.0104)	(0.0107)		(0.00994)	(0.0102)
Male			0.618**			0.657**
			(0.287)			(0.278)
USA empl.			0.408			0.404
			(0.290)			(0.275)
EU empl.			0.223			0.150
			(0.332)			(0.315)
Constant (cut 1)	1.402***	1.271***	2.124***	1.199***	1.070***	1.945***
	(0.196)	(0.221)	(0.434)	(0.185)	(0.211)	(0.417)
Constant (cut 2)	2.305***	2.180***	3.055***	2.038***	1.915***	2.816***
	(0.234)	(0.255)	(0.457)	(0.216)	(0.237)	(0.436)
Observations	488	488	488	488	488	488

Note: The dependent variable in all six specifications is *ShareData*, which takes on three values. An increase in *Ranking* represents a decrease in the ranking of the institution. Thus, a negative coefficient stands for more data availability in higher-ranked institutions. All coefficients are estimated using an ordered probit model. Standard errors are reported in parentheses and the significance levels are reported as follows: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

In specifications (1), (2) and (3), we include the researchers' personal attitude towards sharing as proxied by $Shareothermat$. In specifications (4), (5) and (6), we keep only the structural variables. The results do not change.

Among the structural variables we find two main positive and significant predictors for sharing: $FullProfessor$ and $Ranking$. First, researchers at an advanced stage of their career

tend to share more than junior scholars. The coefficient is positive and highly significant at the 1% level throughout all six specifications. Being full professor increases the likelihood of some sharing (full sharing) as compared to no sharing (some sharing) by 7.2% (1.6%). The career concerns of young researchers might play a role in the decision to share data. Data sharing creates competition as it permits other researchers to use a data set before its creator can fully exploit it in further research. As an additional publication is arguably more valuable in terms of career concerns for (untenured) junior scholars than for full professors, it is not surprising that full professors share their data more frequently.[9] An alternative explanation is that full professors have more financial resources to hire research assistants to create and update the data&code section on their webpages. We control for this by adding the number of research grants that can be thought of as a proxy for the amount of financial resources available to the researcher. In both models, with and without researchers' personal attitudes towards sharing, the sign and significance level of $FullProfessor$ remain unchanged.

Second, researchers affiliated with higher-ranked institutions tend to share more. To illustrate, being affiliated with an institution ranked 50 positions higher increases the likelihood of some sharing (full sharing) as compared to no sharing (some sharing) by 3.3% (0.7%). Controlling for $USAempl$, the significance of $Ranking$ diminishes in specifications (3) and (6) as most of the top-ranked institutions are in the U.S. in the rankings we use in our data set.

The strongest predictor of $ShareData$ is $Shareothermat$. The coefficient is positive and highly significant. Sharing other material increases the likelihood of some sharing (full sharing) as compared to no sharing (some sharing) by 12.2% (3.8%). This result suggests that personal attitudes have the largest impact on the likelihood of data sharing. Apart from $Male$,[10] all other variables are insignificant throughout all specifications.

[9] An intuition for this line of argument is provided by Stanford's tenured Professor Robert Hall (2009). Writing about the career incentives of tenured professors, he states: "Now that you have tenure, the number of papers you produce is amazingly irrelevant."

[10] Three female researchers (out of a total of 97) in our data set (partly) share data. We do not have sufficient data to study gender effects.

4 Conclusion

In this paper, we provide empirical evidence that the current status quo in economics is to not share data or facilitate access to data and codes used for empirical work. We investigate the relation between sharing and observable researchers' characteristics using an ordered probit regression. We find three statistically significant relations across specifications in our data. The likelihood to share is positively associated with sharing other material, being full professor and being affiliated with a higher-ranked institution.

In a recent article, Angrist and Pischke (2010) argue that better research design and the consequent causal interpretation of the regression coefficients "is taking the con out of econometrics". Even though the identification strategy is essential for thorough empirical work, without the possibility of replication and extension provided by sharing data and codes, doubts about the credibility of empirical work remain (Hamermesh, 2006; McCullough at al., 2008). On the one hand, the public availability of data may increase the credibility of empirical work and the reputation of authors as their work might be replicated by others. On the other, it may facilitate new research as both data and codes would be readily available (McCullough and Vinod, 2003).

Acknowledgement: *We gratefully acknowledge financial support from the German Research Foundation (DFG) under the European Data Watch Extended Project. We thank Daniel Hamermesh, Dietmar Harhoff, Fabian Herweg, Mark Schankerman, Olaf Siegert, Ralf Toepfer, Sven Vlaeminck, Gert G. Wagner, Joachim Wagner and Joachim Winter for their helpful comments and suggestions. We also thank participants of the Third LERU Doctoral Summer School "Beyond Open Access: Open Education, Open Data and Open Knowledge" and seminar participants in Munich and Hamburg for their valuable comments as well as Christina Wallner and Jonas Rathfelder for excellent research assistance. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.*

References

[1] 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.

[2] 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, 99-119.

[3] 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, 3-30.

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

[5] Card, D., DellaVigna, S., 2013. Nine facts about top journals in economics. Journal of Economic Literature, forthcoming.

[6] Christie, W., Schultz, P.H., 1994. Why do NASDAQ market makers avoid odd-eighth quotes?. Journal of Finance, 49, 1813-1840.

[7] Coupé, T., 2003. Revealed performances: Worldwide rankings of economists and economics departments, 1990-2000. Journal of the European Economic Association, 1, 1309-1345.

[8] 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, 587-603.

[9] Ferraz, C., Finan, F., 2008. Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes. Quarterly Journal of Economics, 123, 703-745.

[10] Hall, R.E., 2009. Managing your career as an economist after tenure. CSWEP Newsletter, Winter 2009. 4-5.

[11] Hamermesh, D.S., 2006. Viewpoint: Replication in economics. Canadian Journal of Economics, 40, 715-33.

[12] McCullough, B.D., 2009. Open access economics journals and the market for reproducible economic research. Economic Analysis & Policy, 39, 117-126.

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

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

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

[16] Moffitt, R.A., 2007. Research data integrity in economics and other social sciences. Presentation to National Academies Committee on Science, Engineering, and Public Policy, Committee on Assuring the Integrity of Research Data, April, 2007.

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