Data Citation Practices among Taiwan Social Scientists: Some Preliminary Findings

Chi-Shiou Lin[1] and Ching-yi Lai[2]

[1] Associate Professor, Department of Library & Information Science,
National Taiwan University, Taipei, Taiwan
[2] MA Student, Graduate Institute of Library & Information Science,
National Taiwan University, Taipei, Taiwan
chishioulin@ntu.edu.tw

Abstract. This poster describes some preliminary findings from a larger study that investigates the characteristics of social sciences papers that have used externally acquired datasets in the empirical analyses. Specifically, this poster will focus the characteristics of data reporting and data citation behaviors. Using 511 sample articles published within 2011 and 2015, the distributions of data reporting and data citation as well as paper locations where data were mentioned will be described and discussed.

Keywords: Data Citation, Data Reporting, Social Sciences.

1 Introduction

Data sharing and data reuse are becoming common practices in scientific research. As such, a few previous studies have begun to examine data citation practices among researchers, including whether data are cited as references and what dataset attributes are reported and how they are described in those citing papers [1-3]. Appropriate citation behaviors may enhance transparency and accountability of scientific research as well as promote possible reuses of existing data. This study thus examines the current data citation behaviors among Taiwan social scientists so as to understand what can be done to promote a better practice. In this study, *data reporting* and *data citation* are differentiated. The former refers to any textual description of externally-acquired data occurring in various locations of the citing paper, e.g., abstract, main text, tables, and acknowledgements. The latter specifically refers to the acknowledging of the data source in the form of a reference entry. In this poster, we will present some findings based on an analysis of 511 social sciences research articles, published between 2011 and 2015 in Taiwan, that have incorporated external data in their analyses. The sample articles are representative of five subject domains: economics, education, political science, sociology, and psychology.

2 Research Method

The sample articles were manually identified for the systemic content analysis of data citation characteristics. The journal source, identification of sample articles, and the data coding procedures are explained as follow.

Journal Source. The 2015 journal list of *Taiwan Social Science Citation Index* (TSSCI) was used as the basis of journal selection. This citation database selectively indexes research journals of high quality and are representative of Taiwanese scholarship. All journals listed under the five subject categories are carefully examined to identify articles using external data. The numbers of journals varied among the subject categories (Economics: 7 journals; Education: 22; Political Science: 11; Sociology: 11; Psychology: 4).

*Identification of Sample Articles***.** Excluding non-research papers (e.g., editorials, commentaries, book reviews), those journals together published 4207 research papers within the five years. Each research article was manually scanned to ascertain whether it had employed external data in its analysis. Consequently, 511 data-reuse articles were identified (12.15% of the total research articles). But the numbers of data reuse papers in those five subject domains varied greatly (see the following section).

*Data Coding***.** For each article, the number of externally-acquired datasets was recorded. How each dataset is described was further recorded, including whether it was encoded as a bibliographic entry as well as the locations where it was textually described, particularly, the more information-dense locations like the abstract, tables, and acknowledgements. It should be noted that a particular dataset can be cited as a reference entry and, at the same time, be textually reported in multiple locations of the citing paper. All those situations were recorded for subsequent analyses.

Based on the previous procedures, 875 datasets used in the 511 articles were identified. The distributions of the articles and the reused datasets are as in Table 1.

Table 1. The distributions of data reuse articles and the reused datasets (2011-2015)

Subject	N. of papers	N. of datasets	Avg. datasets (per paper)	St. D.
Economics	215	452	2.10	1.58
Education	105	124	1.18	0.62
Political Science	93	157	1.69	1.25
Sociology	81	120	1.48	1.28
Psychology	17	22	1.29	0.69
Total	**511**	**875**	**1.71**	**1.35**

3 Preliminary Findings

3.1 Data Reporting and Data Citation within the Five Subjects

Table 2 shows how the articles report or cite data. 39.73% of the articles did cite data in the references. Proportionally, political scientists were the best at citing data, followed by education researchers and sociologists. Economists, being the heaviest users or external data in our sample, were the worst in data citation. Further, observing the paper locations where data reporting occurs, economists also rarely report data in the more information-dense locations, e.g., abstract, acknowledgements, or table notes. Researchers of political science, education, and sociology had made better uses of the abstracts and table notes to accredit the external datasets.

Table 2. Data reporting and data citation in the Five Subjects

	Data Reporting								Data Citation	
	Main text		**Abstract**		**Acknow.**		**Tables**		**References**	
	n.	**%**	**n.**	**%**	**n.**	**%**	**n.**	**%**	**n.**	**%**
Eco. (N=215)	215	100.00	35	16.28	6	2.79	45	20.93	58	26.98
Edu. (N=105)	105	100.00	81	77.14	4	3.81	8	7.62	52	49.52
Pol. (N=93)	93	100.00	42	45.16	22	23.66	54	58.06	55	59.14
Soc. (N=81)	81	100.00	62	76.54	14	17.28	12	14.81	32	39.51
Psy. (N=17)	17	100.00	13	76.47	3	17.65	2	11.76	6	35.29
Total(N=511)	**511**	**100.00**	**233**	**45.60**	**49**	**9.59**	**121**	**23.68**	**203**	**39.73**

3.2 Co-Presences of Data Description in a Paper

365 of the 511 papers (71.43%) had mentioned data beyond in the main text. Data description in the information-dense locations such as the abstracts, tables, acknowledgements, and bibliographic references helps raise data visibility. 188 articles (36.79%) had mentioned data in one additional location outside the main text; 124 articles (24.27%) had mentioned data in two additional locations.

Table 3 shows that, when an article mentioned data in the more information-dense location outside the main text, the abstract is usually the preferred location, followed by the references. Here we also see a great improvement of economics authors' data citation behavior. When an economics author began to mention data outside the main text, there's a greater chance that he/she might cited it as a reference entry.

Table 4 shows that, when an article mentioned data in two additional locations beyond the main text, the overall percentage of data citation became very high (79.84% of the 124 articles). And for the economics researchers whose data citation behaviors were generally poor, the percentage of data citation could rise up to 85.71%. This suggests that scholars who have stronger tendency in emphasizing and accrediting external data also have stronger tendency in conducting data citation.

Table 3. Distribution of data mentions in the main text plus one additional location

	Data Reporting						Data Citation	
	Main+Abs.		**Main+Ack.**		**Main+Tables**		**Main+Ref.**	
	n.	**%**	**n.**	**%**	**n.**	**%**	**n.**	**%**
Eco. (N=62)	17	27.42	2	3.23	18	29.03	25	40.32
Edu. (N=46)	37	80.43	0	0.00	0	0.00	9	19.57
Pol. (N=28)	11	39.29	1	3.57	7	25.00	9	32.14
Soc. (N=43)	32	74.42	1	2.33	2	4.65	8	18.60
Psy. (N=9)	7	77.78	0	0.00	0	0.00	2	22.22
Total (N=188)	**104**	**55.32**	**4**	**2.13**	**27**	**14.36**	**53**	**28.19**

Table 4. Distribution of data mentions in the main text plus two additional locations

	Data Reporting				Data Citation			
	Main+Abs.+ Tables		**Main+Abs.+ Acknowl.**		**Main+Abs.+ References**		**Main+Tables+ References**	
	n.	**%**	**n.**	**%**	**n.**	**%**	**n.**	**%**
Eco. (N=35)	2	5.71	3	8.57	9	25.71	21	60.00
Edu. (N=34)	0	0.00	1	2.94	33	97.06	0	0.00
Pol. (N=29)	6	20.69	1	3.45	3	10.34	16	55.17
Soc. (N=20)	3	15.00	3	15.00	13	65.00	1	5.00
Psy. (N=6)	0	0.00	3	50.00	2	33.33	1	16.67
Total (N=124)	**11**	**8.87**	**11**	**8.87**	**60**	**48.39**	**39**	**31.45**

4 Temporary Conclusion

This poster presents on the distributions of data reporting and data citation behaviors among Taiwan social scientists. The percentage of papers engaging in data citation as well as papers with multiple data description locations are described and discussed. Our future analyses will continue to examine the quality of information about those data. We will further examine what information is offered about the datasets (i.e., the creator/disseminator, data title, publishing year, DOI/URL, other access information) so as to understand whether sufficient information has been provided about the reused data and what can be done to improve future data citation quality and accuracy.

References

1. Mooney, H.: Citing data sources in the social sciences. Learned Publishing 24(2), 99–108 (2011).
2. Piwowar, H., Vision, T., Carlson, J.: Beginning to track 1000 datasets from public repositories into the published literature. Proceedings of the American Society for Information Science and Technology 48(1), 1–4 (2011).
3. Henderson, T.: Data citation practices in the CRAWDAD Wireless Network Data Archives. D-Lib Magazine 21(1/2) (2015).