KHAZANAH
RESEARCH
INSTITUTE

WORKING PAPER 4/20 | 17 DECEMBER 2020

Open Government Data for Academic Research

Ashraf Shaharudin

Khazanah Research Institute

The **KRI Working Papers** are a series of research documents by the author(s) containing preliminary findings from ongoing research work. They are a work in progress published to elicit comments and encourage debate. In that respect, readers are encouraged to submit their comments directly to the authors.

The views and opinions expressed are those of the author and may not necessarily represent the official views of KRI. All errors remain the authors' own.

WORKING PAPER 4/20 | 17 DECEMBER 2020

Open Government Data for Academic Research

This working paper was prepared by Ashraf Shaharudin from the Khazanah Research Institute with research assistance from Shariman Arif Mohamad Yusof and Amos Tong Huai En. The author is grateful for the valuable comments from Rachel Gong, Gregory Ho Wai Son, Siti Aiysyah Tumin, Mohd Amirul Rafiq Abu Rahim and Puteri Marjan Megat Muzafar.

Authors' email address: **Ashraf.Shaharudin@KRInstitute.org / ashrafshaharudin@gmail.com (after 6 Dec 2021)**

Information on Khazanah Research Institute publications and digital products can be found at **www.KRInstitute.org**.

This is the second paper in the open government data series, as part of the Networked Nation project that looks into digital issues in Malaysia. The first paper, Open Government Data: Principles, Benefits and Evaluations, provides a general overview of the open government data.

Executive Summary

- Open government data empowers researchers to conduct meaningful analysis, especially to address important issues in society. It is also a means to achieve open science, which aims to promote a more accurate verification of scientific findings via peer-review and replication, and reduce duplication in collecting data.

- This paper provides insights into the perception of open government data among researchers who do research on Malaysia and use open government data based on an online survey.

- The level of open government data in Malaysia is generally considered unsatisfactory among most respondents. Most researchers find data provided by the Malaysian government falls short in three areas: completeness, granularity and timeliness.

- The survey also highlights the low favourability of Malaysia's central open government data platform, data.gov.my portal, compared to other open data platforms. It is the last go-to source of data among respondents and has the highest percentage of respondents who considered the platform not useful.

Table of Contents

1. Introduction

1.1. Objective

This paper provides insights into the perception of open government data among academic researchers who conduct research on Malaysia and use open government data. These insights were obtained through an online survey.

The outline of this paper is as follows. This section provides the background of open government data and academic research. Section 2 describes the methods employed. Section 3 provides the findings of the survey. Lastly, Section 4 discusses the findings and Section 5 concludes.

1.2. Background: Open government data

Data is a valuable yet non-exhaustible[1] commodity. In research production, the deployment of capital (e.g. to purchase lab equipment) and labour (e.g. researcher) for a research project will diminish their availability for other projects. However, this is not the case for data, as its consumption by a group will not reduce its availability for others.

In contrast to the "tragedy of the commons" where consumption of common-pool resources[2], such as forest resources, eventually reduces their availability, data generates more value with greater use[3]. This means the more data that is made open, the more value can be generated. In the context of research production, **open data allows society to harness collective intelligence in generating knowledge[4].**

The focus of this paper lies in open government data. Government data refers to data collected by the public sector through various means including census, survey, private sector reporting as well as through technology such as weather facility and satellite imagery. Since this data is collected and held using public funds, it is imperative to maximise the potential (public) value[5] of this data including in knowledge generation to advance society.

Open government data is one of the Malaysian government's focus areas as stipulated in the Department of Statistics Malaysia (DOSM) Transformation Plan 2015 – 2020, the Public Sector Information and Communication Technology (ICT) Strategic Plan 2016 – 2020 and the Communications & Multimedia Blueprint 2018 – 2025.

---

[1] Non-exhaustibility is different from non-excludability. Data is non-exhaustible, but its excludability depends on who can be excluded from accessing it. Publicly accessible data is non-excludable (i.e. no one can be excluded from accessing it) whereas privately accessible data is excludable.

[2] in a non-sustainable way

[3] A scenario termed "the comedy of the commons" by Rose (1986)

[4] Janssen, Charalabidis, and Zuiderwijk (2012) & Ryan, Gambrell, and Noveck (2020)

[5] Data can be used to generate either private or public value. The challenge is to reap the most public value possible from public data while also allowing public data to be used for private gain.

Data is considered open when it is free from legal and technical constraints to be used by anyone at any time from anywhere[6]. Open data should fulfil six salient features: completeness, granularity, timeliness, accessibility, machine-processability and non-proprietorship (Table 1.1). This means, data that is merely published online without fulfilling the set of features listed does not constitute open data as it is not accessible for consumption in a full sense.

Table 1.1: Salient features of open government data

Feature	Aspects
1. **Completeness**	• All data is **open by default**, unless with valid justifications for closure such as privacy and security concerns. • **Permanence**, i.e. data available online should remain online. • Comprehensive **metadata** included. • Initiative to **digitise non-electronic data** such as physical artefacts is encouraged.
2. **Granularity**	• Highest possible level of **granularity**. • If possible, data are provided in its **original and unmodified form**.
3. **Timeliness**	• Data is made available as **quickly** as possible.
4. **Accessibility**	• **Open license**. • **Free** of charge. • **Downloadable** via the Internet. • **No legal/technical barrier**.
5. **Machine-processability**	• Data in a form **readily processable by a computer**
6. **Non-proprietorship**	• Data is available in **a format in which no entity has exclusive control**.

Source: International Open Data Charter (n.d.), Sunlight Foundation (2010) & Open Knowledge Foundation (n.d.) as summarised in Ashraf (2020b)

Open data is one way of achieving open science. The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines open science as "the movement to make scientific research and data accessible to all"[7]. The objectives of open science include allowing a more accurate verification of scientific findings through peer-review and replication, reducing duplication in collecting data and promoting citizens' engagement in science[8]. The Covid-19 pandemic demonstrates the importance of open science to foster scientific collaboration and accelerate public health response[9]. While open science requires efforts from various stakeholders especially from the research community, governments have a big role in realising open science including by making government data open.

Malaysia did not perform well in multiple established global open government data evaluations. In the latest Open Data Barometer (ODB), Global Open Data Index (GODI), Open Data Inventory (ODIN) and Open Budget Index (OBI), Malaysia was behind the Philippines, Singapore and Indonesia, as well as many other developing countries (Table 1.2). Malaysia ranked eighth from the bottom out of 94 places in the latest GODI.

[6] International Open Data Charter (n.d.) & Open Knowledge Foundation (n.d.)
[7] UNESCO (2017)
[8] UNESCO (2017) & OECD (n.d.; 2015).
[9] See Ashraf (2020a) on the role of open science in responding to the Covid-19 crisis.

Table 1.2: Open government data evaluations ranking, selected countries

ODB 2016 (/115)		GODI 2016/17 (/94)		ODIN 2018/19 (/178)		OBI 2019 (/118)	
Mexico	11	Brazil	8	Singapore	1	South Africa	2
Brazil	18	Mexico	11	Mexico	22	Mexico	4
Philippines	22	Singapore	17	Philippines	41	Brazil	6
Singapore	23	India	32	Indonesia	49	Philippines	10
India	33	South Africa	43	India	55	Indonesia	18
Indonesia	38	Turkey	45	Brazil	56	Thailand	30
Turkey	40	Thailand	51	South Africa	65	Turkey	46
South Africa	46	Philippines	53	**Malaysia**	**69**	India	53
Malaysia	**53**	Indonesia	61	Turkey	74	**Malaysia**	**55**
Thailand	53	**Malaysia**	**87**	Thailand	126	Singapore	N/A

Source: World Wide Web Foundation (2017), Open Knowledge Foundation (2016), Open Data Watch (2019), International Budget Partnership (2020) as compiled in Ashraf (2020b)

1.3. Background: Research production in Malaysia

Academic researchers produce various types of output including patents, policy reports and media communication. However, citeable academic publication such as a journal article is the major and easily measurable output of academic researchers. Based on the Scopus database, Malaysia's yearly published citable[10] research documents, i.e. articles, reviews and conference papers, have increased from 11,227 in 2009 to 35,111 in 2019. Other neighbouring countries, namely Singapore, Thailand, Indonesia and Vietnam also have a similar trend (Figure 1.1). The growth of citable documents for Malaysia started to accelerate in 2007.

The number of citable documents is a function of a country's size (which determines the number of researchers, research institutions, research expenditure, etc.). **Even though Malaysia has a higher number of citable documents compared to Singapore and Thailand, the country has a lower h-index** (Figure 1.2). A country's h-index is the country's number of academic documents (*h*) that have received at least *h* citations. It reflects the quality of research publications.

There is a positive and significant relationship between the level of open government data and h-index as discussed in Box 1.

Figure 1.1: Citable documents, selected southeast Asian countries, 1996 – 2019

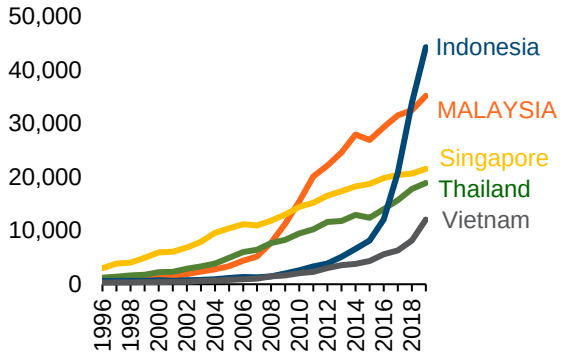
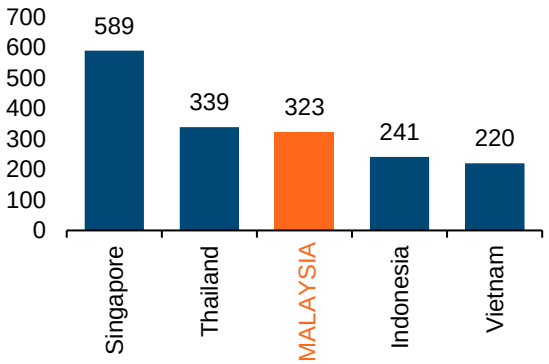


Figure 1.2: H-index, selected southeast Asian countries



Source: Scimago Lab (2020b)

[10] Note that citable documents are not cited documents, but simply documents that can be cited.

Box 1: Does open government data improve academic publication?

A cross-sectional analysis was conducted to investigate the relationship between open government data and academic output. Detailed methods, findings and discussion are presented in **Appendix A**.

This study uses economic tools in modelling research production. A simple cross-country research production model is introduced. Suppose there are two factors[11] in research production, namely scientific labour (i.e. researcher) and capital investment (i.e. research expenditure excluding labour costs). Consider these two factors as non-substitutable, which means an increase in one factor without a corresponding increase in the other factor will not increase the output (i.e. the academic publication).

Two drivers of research production are considered in this study: the country's economic development (*dev*) and open data (*od*) level. The selection of economic development as another driver of concern in this study, apart from the open data level, is because it may capture a lot of cross-country variations including the quality of researchers and pre-existing research infrastructure.

The final estimation model is written as equation below (derivation in Appendix A).

$$\ln(Q) = \tau + \gamma \ln(mK + c + L) + \mu_1 \ln(dev) + \mu_2 \ln(od)$$

where Q: Research output
K: Research capital
L: Researchers
m: Gradient between L and K
γ and μ: Rate of returns
c : L when K = 0

Analysis in this study is conducted for three samples: research in all sectors (sample 1), research in the higher education sector (sample 2) and social science research in the higher education sector (sample 3).

Q is represented by H-index whereas *od* by Open Data Barometer (ODB) score. H-index for all field and social science versus ODB score are plotted in Figure 1.3 and 1.4 respectively.

Figure 1.3: H-index (all field) versus Open Data Barometer (OBD) score, 2016

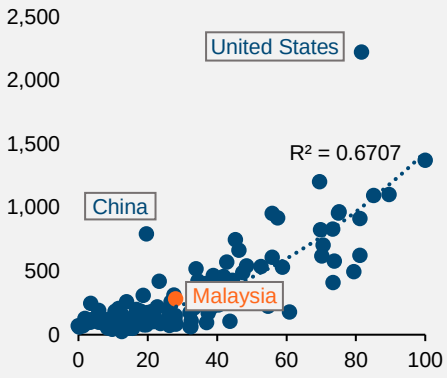
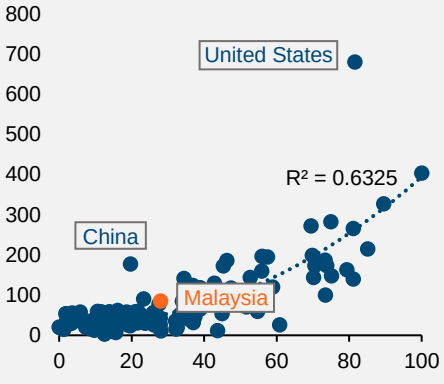


Figure 1.4: H-index (social science) versus Open Data Barometer (OBD) score, 2016



Source: World Wide Web Foundation (2017) & Scimago Lab (2020a)

Table 1.3 presents the estimation results. To mitigate the heteroskedasticity issue, robust standard errors are reported. The Shapiro-Wilks normality test ran found that the residuals for sample 2 are not normally distributed. Nevertheless, there is no multicollinearity issue among variables, as verified through the variance inflation factor (VIF) test.

Table 1.3: Estimated effects of research production factors and drivers on research output

	Sample 1		Sample 2		Sample 3	
	(A)	(B)	(C)	(D)	(E)	(F)
Coefficients of:						
$\ln(mK + c + L)$	2.8503*** (0.2113)	2.7862*** (0.2039)	2.8393*** (0.2246)	2.7410*** (0.2247)	3.4795*** (0.4318)	3.4629*** (0.4238)
$\ln(dev)$	0.1254*** (0.0370)	0.1070*** (0.0374)	0.1106** (0.0524)	0.0825* (0.0480)	0.0413 (0.0745)	0.0150 (0.0717)
$\ln(od)$	**0.2377*** (0.0775)**	**0.2024*** (0.0735)**	**0.2666** (0.1351)**	**0.2238* (0.1311)**	**0.2937** (0.1391)**	**0.2050 (0.1498)**
eng	-	0.1011*** (0.0373)	-	0.1429*** (0.0539)	-	0.1390** (0.0686)
Intercept (τ)	-4.8339*** (0.5072)	-4.4420*** (0.4933)	-4.5994*** (0.6277)	-4.0354*** (0.6833)	-6.7560*** (0.9334)	-6.3077*** (0.8948)
Sample size	83	83	76	76	51	51
R^2	0.8933	0.9001	0.8552	0.8680	0.8577	0.8659

Note: Asterisks *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Robust standard errors are in parentheses.

Open data is found to be statistically significant with a positive coefficient for all three samples except with the inclusion of *eng* variable in sample 3. However, economic development is also found to be statistically insignificant in sample 3. This may be because the size of sample 3 is small and for countries in the sample, research output is almost entirely accounted for by the factors of production, namely scientific labour and capital investment.

Undoubtedly, researchers and capital investment are the factors of research, hence they are the main determinants of research output. Rather interestingly, an improvement in the GDP per capita and the ODB score would render a comparable improvement in the h-index. This highlights the significance of open government data for academic research.

[11] In economic terms, 'factors' can be loosely defined as 'input'. Think of them as the necessary items to produce something.

2. Methods

An online survey was conducted in May and June 2020 (2 months) on the perception of open government data among researchers, based locally in Malaysia or abroad, who do research on Malaysia. The survey was disseminated through three ways: (i) by sending out direct emails (twice) to 462 social science academics in public and private universities based in Malaysia, (ii) by reaching out to the administration of 12 public universities to seek help distributing the survey to postgraduate students, and (iii) by publicising the survey through Khazanah Research Institute's social media platforms (Twitter, Facebook, LinkedIn and Instagram).

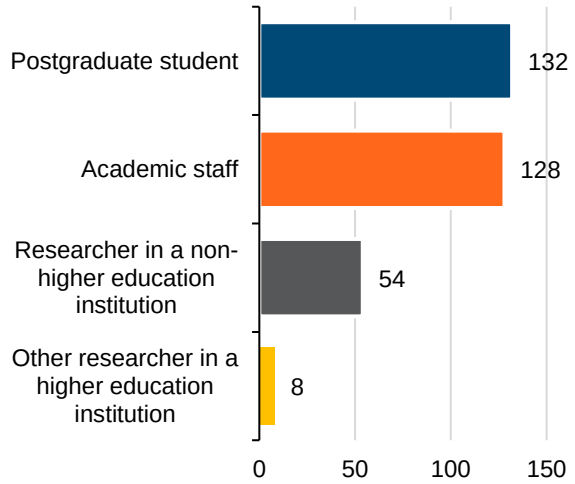
There are three parts in the survey. Responses from the first and second part are used to filter and analyse responses in the third part. The first part establishes the profile of the respondent, the second part evaluates the respondent's familiarity and use of open data and the third part captures the respondent's perception of open government data. Questions in the third part are guided by the features of open data as outlined by the Sunlight Foundation, the International Open Data Charter and the Open Knowledge Foundation and summarised in Table 1.1 in Section 1.2[12]. The survey underwent multiple iterations of pre-testing. The survey questionnaire is presented in Appendix B and the coded responses datasheet is provided along with this paper.

3. Results

3.1. Profile of respondents

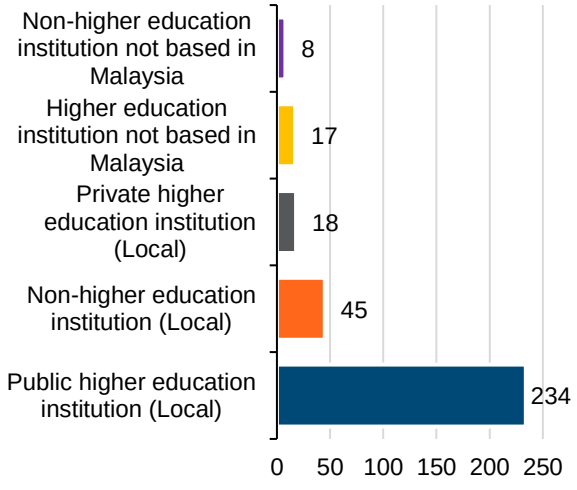
There were 322 responses collected. The sample is skewed towards postgraduate students and academic staff (Figure 3.1) from public higher education institutions (Figure 3.2) majoring in economics (Figure 3.3) with more than six years of research experience (Figure 3.4). The age of respondents is presented in Figure 3.5.

Figure 3.1: Professional role of respondents (frequency)



Source: Online survey conducted by author

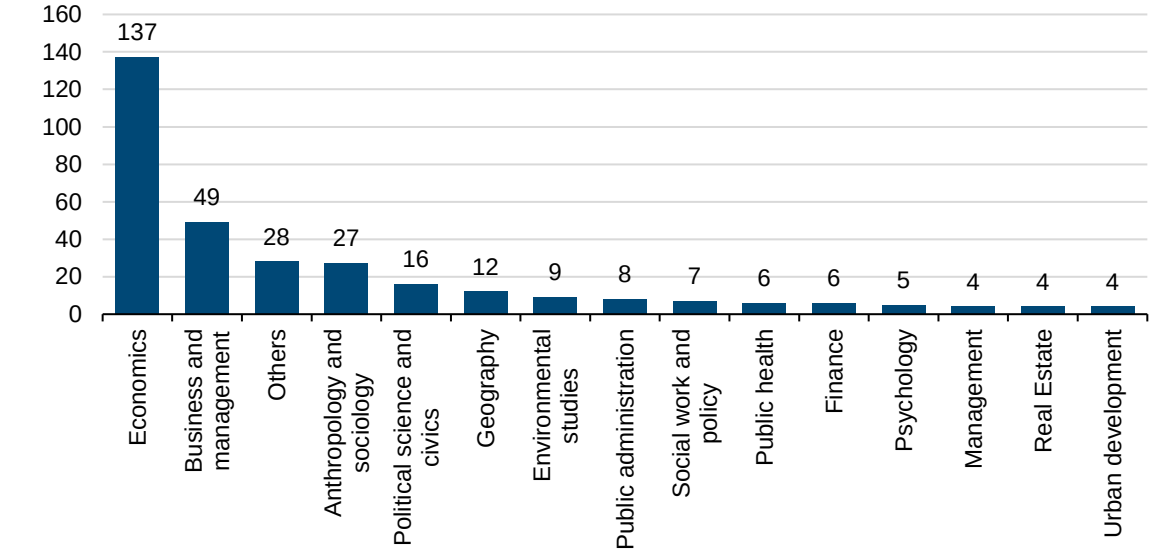
Figure 3.2: Main affiliation of respondents (frequency)



Source: Online survey conducted by author

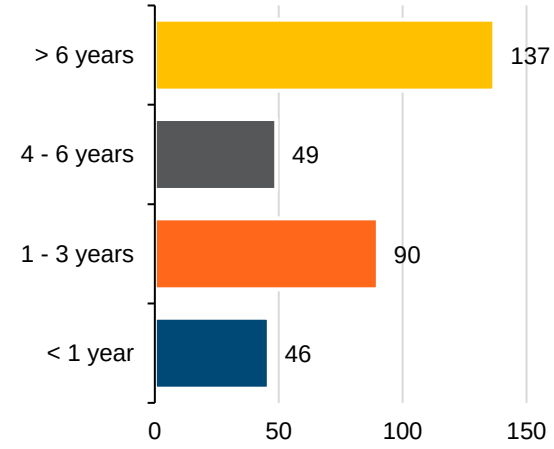
[12] Refer to Ashraf (2020b) for greater discussion on the salient features of open government data

Figure 3.3: Research major of respondents (frequency)



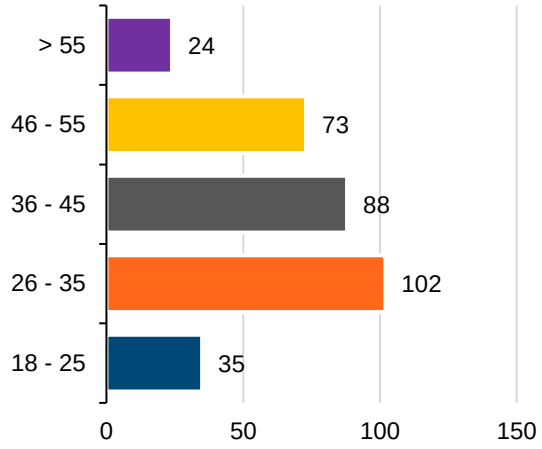
Source: Online survey conducted by author

Figure 3.4: Years of research experience of respondents (frequency)



Source: Online survey conducted by author

Figure 3.5: Age range of respondents (frequency)

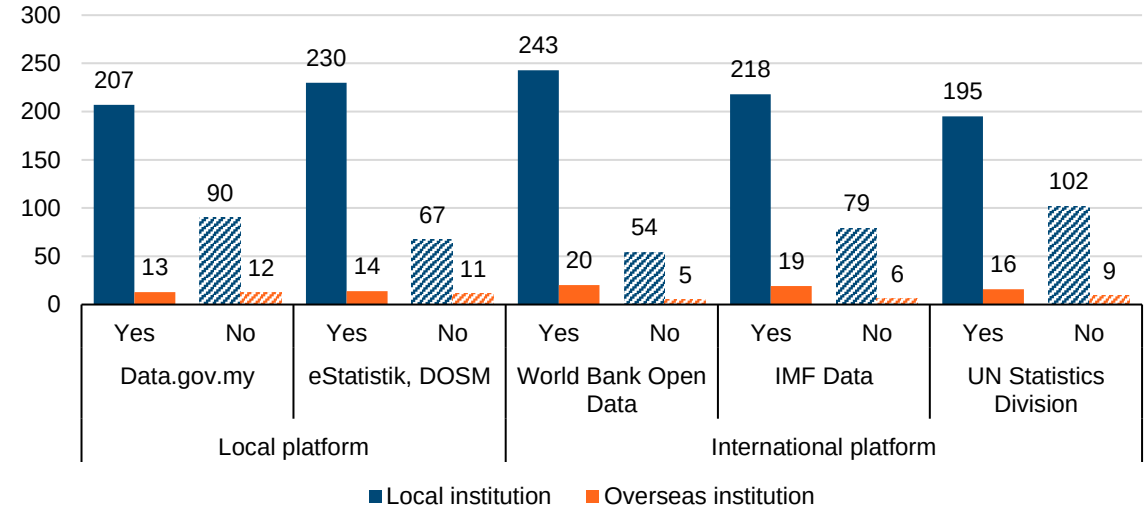


Source: Online survey conducted by author

3.2. Open data familiarity and use

The majority of respondents are familiar with common open data platforms (Figure 3.6). However, some platforms are more popular than others. The World Bank Open Data portal and the Department of Statistics Malaysia (DOSM)'s eStatistik are the two most popular open data platforms. **The two least popular platforms are data.gov.my portal (a one-stop Malaysia's government data platform) and the UN Statistics Division data portal.**

Figure 3.6: Have you heard of these data platforms? (frequency by the locality of institution)

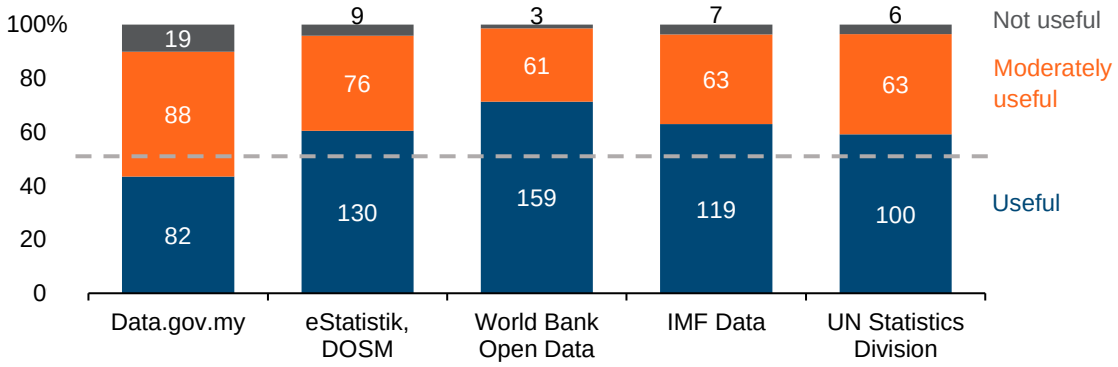


Source: Online survey conducted by author

To analyse the usefulness of different open data platforms, responses from respondents who have (i) never heard (answered 'No' in Figure 3.6) or (ii) never used[13] each of the platforms or (iii) considered more than eight out of the fifteen categories of data in Figure 3.8 as not relevant for them are removed. This is to comb out responses by respondents who are not familiar with each of the open data platform and who may not use open data regularly.

Based on the survey, the World Bank Open Data portal is not only the most popular platform but has the largest share of respondents who considered it to be useful (71%). On the other hand, **10% of respondents who are familiar with the data.gov.my found the platform not useful—the largest share of 'not useful' response compared to other platforms** (Figure 3.7).

Figure 3.7: How useful are these data platforms? (frequency and percentage)

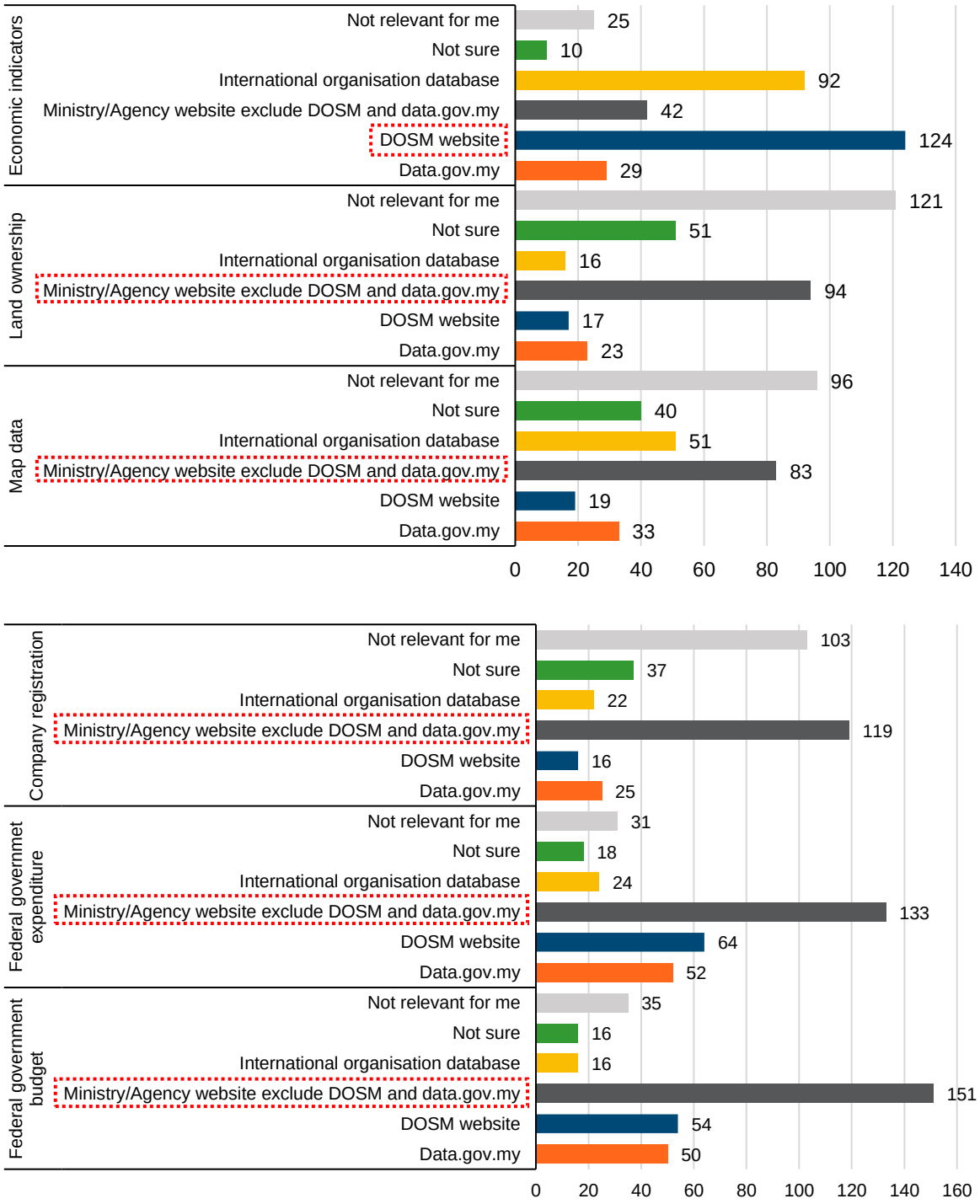


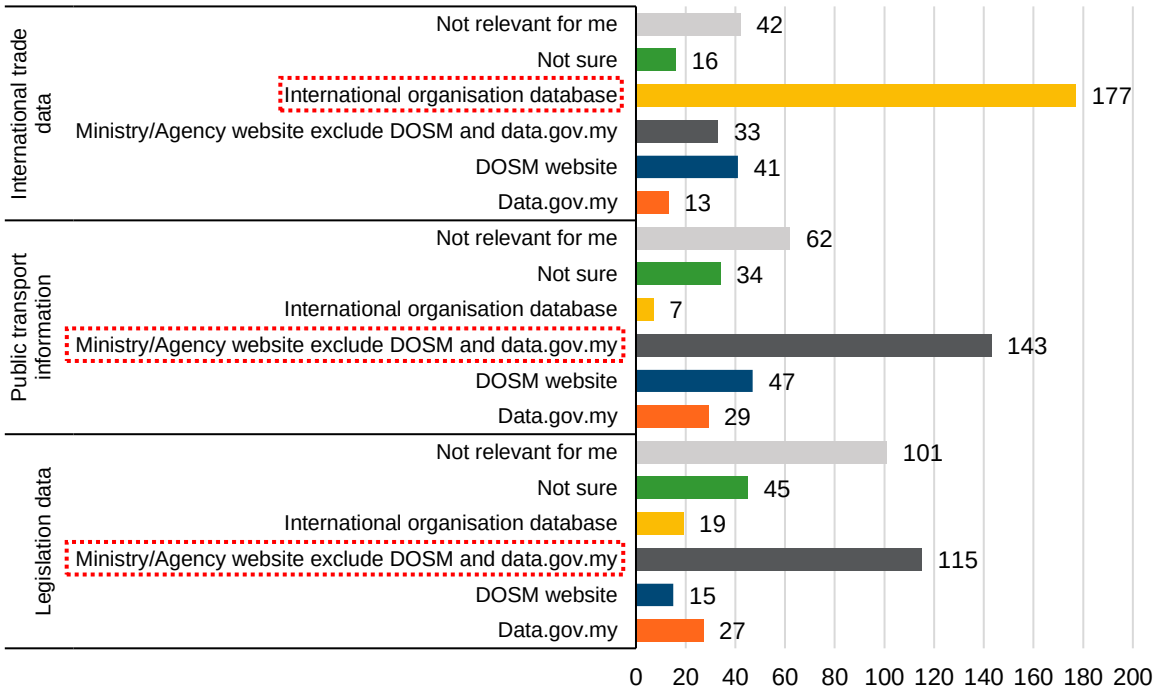
Note: Labels indicate the frequency of responses. The dotted line is the 50% mark.
Source: Online survey conducted by author

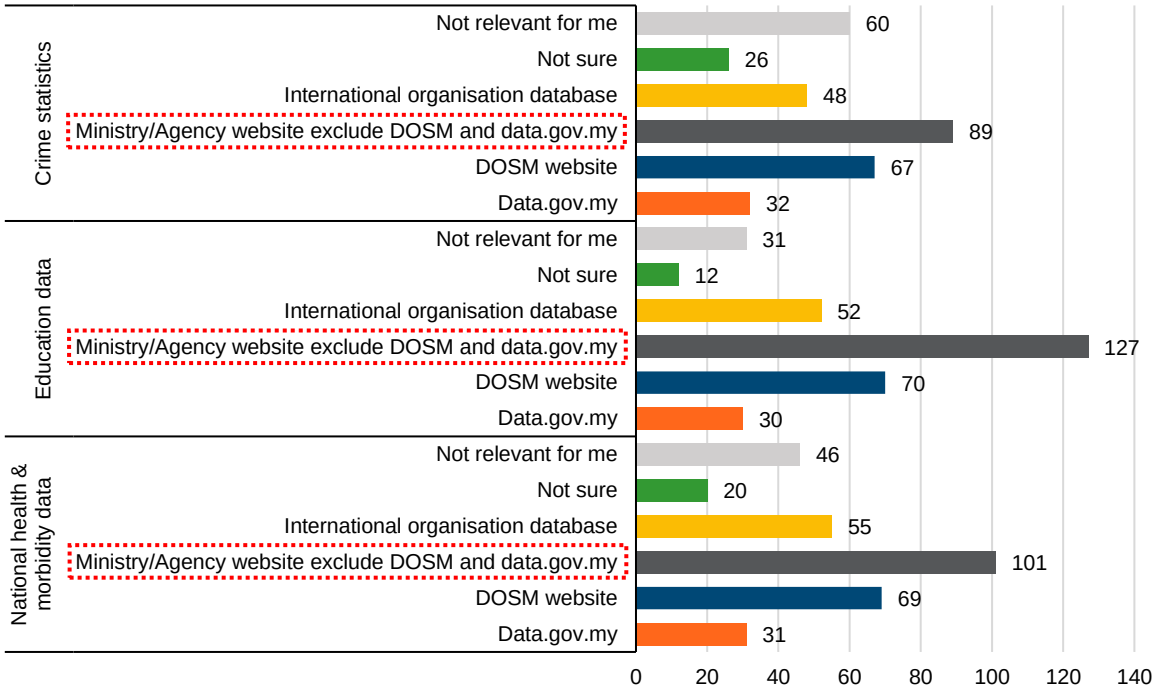
[13] In the survey (refer to Appendix B), respondents were asked to indicate whether they find each open data platform 'Useful', 'Moderately useful', 'Not useful' or 'Never used'. One would expect that respondents who answered that they have not heard a particular platform would answer that they have not used the platform, but this was (weirdly) not always the case. To account for this technical inconsistency, Figure 3.7 eliminate responses of those who have never heard **or** never used each of the respective platform.

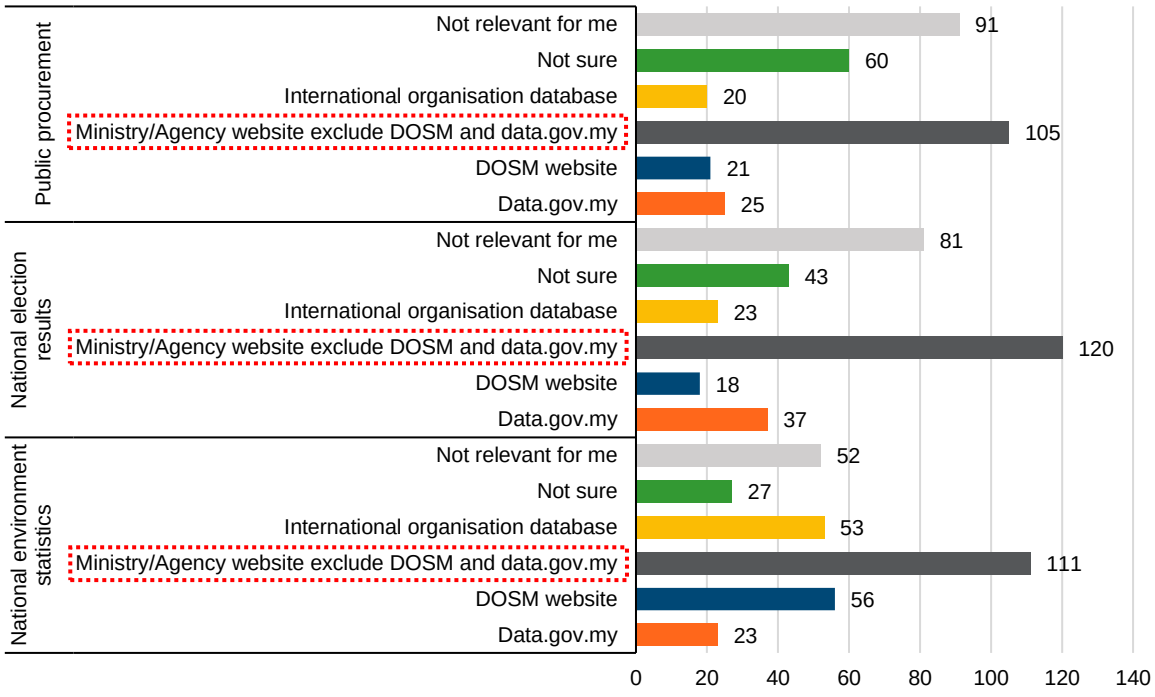
Different platforms are popular for different categories of data (Figure 3.8). For most data categories, respondents would first look for data from the relevant ministry/agency websites, except for economic indicators and international trade data. Most respondents would go to the DOSM's eStatistik for economic indicators and international organisation databases for international trade data. **It is worth noting that even though data.gov.my is supposed to be Malaysia's central open data platform, it is not the first go-to platform for any of the data categories.**

Figure 3.8: Which platform would you go first to obtain these data? (frequency)




International trade data
Not relevant for me 42
Not sure 16
International organisation database 177
Ministry/Agency website exclude DOSM and data.gov.my 33
DOSM website 41
Data.gov.my 13
Public transport information
Not relevant for me 62
Not sure 34
International organisation database 7
Ministry/Agency website exclude DOSM and data.gov.my 143
DOSM website 47
Data.gov.my 29
Legislation data
Not relevant for me 101
Not sure 45
International organisation database 19
Ministry/Agency website exclude DOSM and data.gov.my 115
DOSM website 15
Data.gov.my 27
0 20 40 60 80 100 120 140 160 180 200

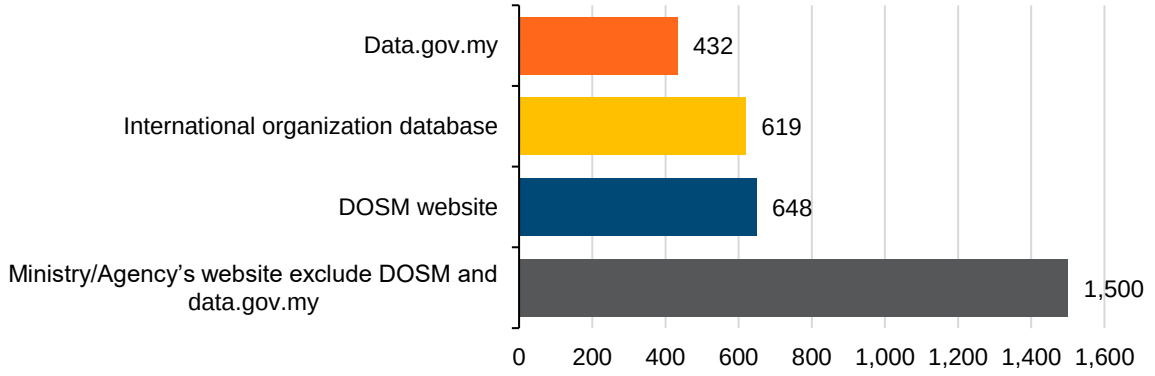

Crime statistics
Not relevant for me 60
Not sure 26
International organisation database 48
Ministry/Agency website exclude DOSM and data.gov.my 89
DOSM website 67
Data.gov.my 32
Education data
Not relevant for me 31
Not sure 12
International organisation database 52
Ministry/Agency website exclude DOSM and data.gov.my 127
DOSM website 70
Data.gov.my 30
National health & morbidity data
Not relevant for me 46
Not sure 20
International organisation database 55
Ministry/Agency website exclude DOSM and data.gov.my 101
DOSM website 69
Data.gov.my 31
0 20 40 60 80 100 120 140



Source: Online survey conducted by author

Figure 3.9 shows the aggregated preferred data source of respondents by summing up their responses across all data categories. In aggregate, ministry/agencies' websites are the most popular source of data, followed by DOSM's website. **Data.gov.my portal is the least popular data source** (Figure 3.9). Recall that data.gov.my also has the largest share of respondents who found the platform not useful.

Figure 3.9: Aggregate of data source popularity (frequency)



Source: Online survey conducted by author

3.3. Perception of open government data

The perception of open data provided by the Malaysian government through platforms such as data.gov.my and agency/ministry websites among researchers is presented in this sub-section. Responses from respondents who indicated that they have never heard any of the data platforms in Figure 3.6 or indicated that more than eight out of fifteen data categories in Figure 3.8 are not relevant for them are removed. This is to filter responses from respondents who may not be familiar with open government data or may not use open government data regularly. Consequently, 48 responses out of the total 322 were removed (274 remaining).

Figure 3.10 presents the perception of different open data criteria, grouped by the salient features of open government data as described in Table 1.1 in Section 1.2, namely, completeness, granularity, timeliness, accessibility, machine-processability and non-proprietorship as well as additional features, namely, no registration, comparability, usefulness and reliability. Table 3.1 and 3.2 list out statements with the five highest frequencies of 'Agree' and 'Disagree' responses respectively.

Table 3.1: Statements with the five highest frequencies of 'Agree' responses

Statement	Feature	% 'Agree'
I trust the accuracy of most data obtained through government platforms	Reliability	70%
Most data provided online is very useful for me to do my research	Usefulness	66%
Most data is provided in a machine-processable format, such as Excel or CSV and not PDF	Machine-processability	62%
I do not have to purchase or download specific software to obtain most data	Non-proprietorship	61%
Most data is made open free of charge	Accessibility	59%

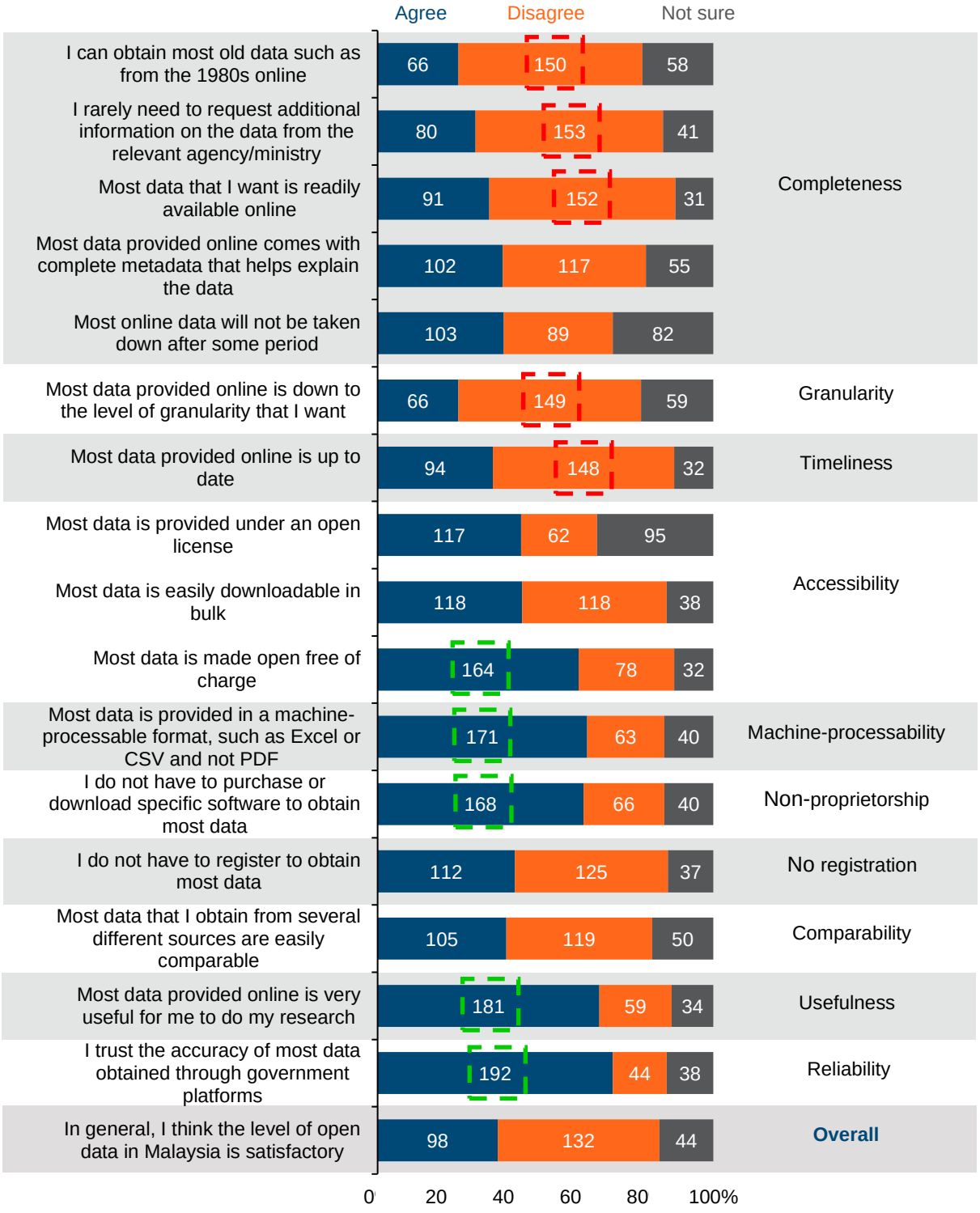
Source: Online survey conducted by author

Table 3.2: Statements with the five highest frequencies of 'Disagree' responses

Statement	Feature	% 'Disagree'
I rarely need to request additional information on the data from the relevant agency/ministry	Completeness	56%
Most data that I want is readily available online	Completeness	55%
I can obtain most old data such as from the 1980s online	Completeness	55%
Most data provided online is down to the level of granularity that I want	Granularity	54%
Most data provided online is up to date	Timeliness	54%

Source: Online survey conducted by author

Figure 3.10: Perception of different open data criteria (frequency and percentage)

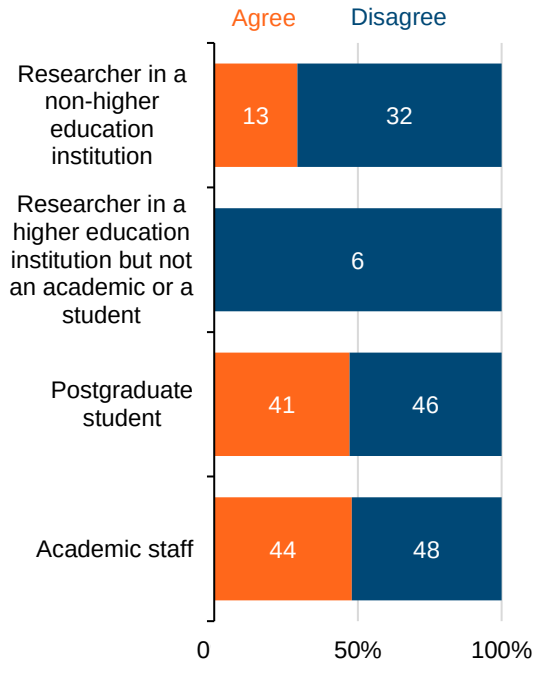


Note: Labels indicate the frequency of responses. Red boxes indicate five responses with the highest 'Disagree' frequencies whereas green boxes indicate five responses with the highest 'Agree' frequencies.
Source: Online survey conducted by author

Most respondents trust the accuracy of government data and they consider them very useful for their research. However, the data provided is not enough for them. Additional information related to the data is usually needed from the relevant agency/ministry. Most data that respondents want is not readily available online or not down to the level of granularity desired. Besides, according to most respondents, both old data such as from the 1980s as well as the most recent data is mostly unavailable online. **In short, according to the survey, open data provided by the Malaysian government largely falls short in three areas, which are completeness, granularity and timeliness.**

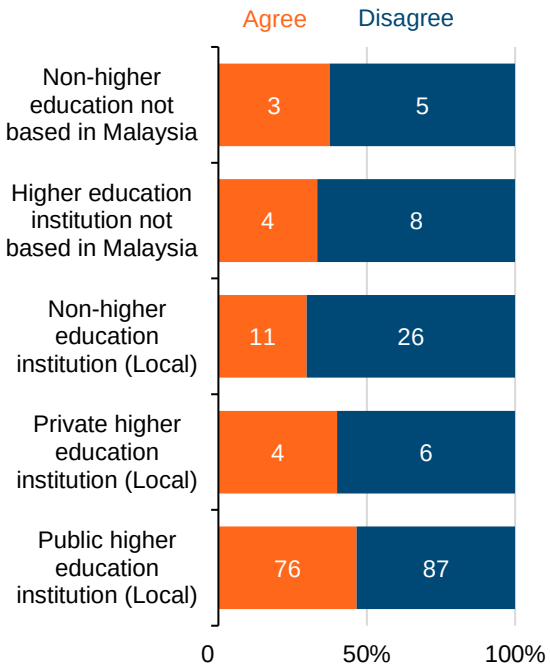
Only 98 out of 274 respondents (36%) reported that they generally think the level of open data in Malaysia is satisfactory (Figure 3.10). To consider sampling skewness, analysis based on profile sub-groups is carried out. 'Not sure' responses are removed in the analysis to eliminate uncertainty. By professional role and main affiliation, less than half of respondents from each sub-group found the level of open data satisfactory (Figure 3.11 and 3.12). By respondents' age and research experience, only one sub-group in each category that has more than half of the respondents indicated that they are satisfied with the level of open data in Malaysia (Figure 3.13 and 3.14). The response is, however, a bit mixed by research major (Figure 3.15).

Figure 3.11: Response to 'In general, I think the level of open data in Malaysia is satisfactory' by professional role



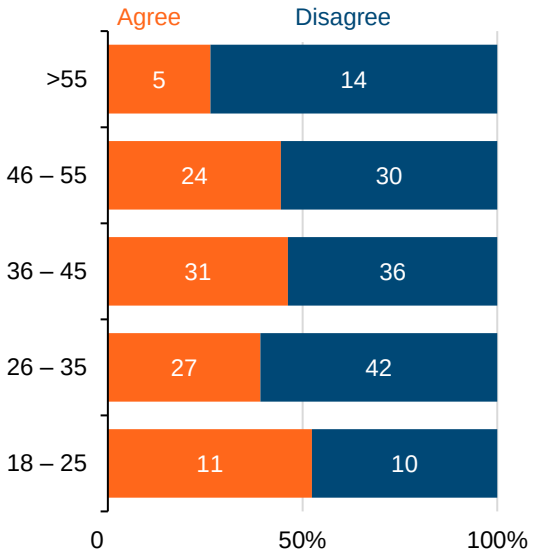
Note: Labels indicate the frequency of responses
Source: Online survey conducted by author

Figure 3.12: Response to 'In general, I think the level of open data in Malaysia is satisfactory' by affiliation



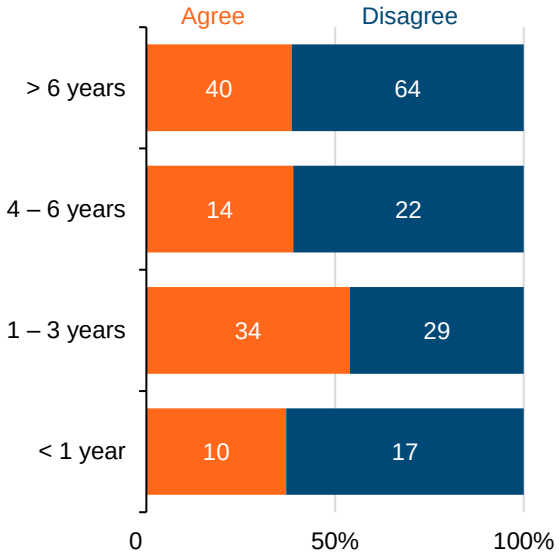
Note: Labels indicate the frequency of responses
Source: Online survey conducted by author

Figure 3.13: Response to 'In general, I think the level of open data in Malaysia is satisfactory' by age range



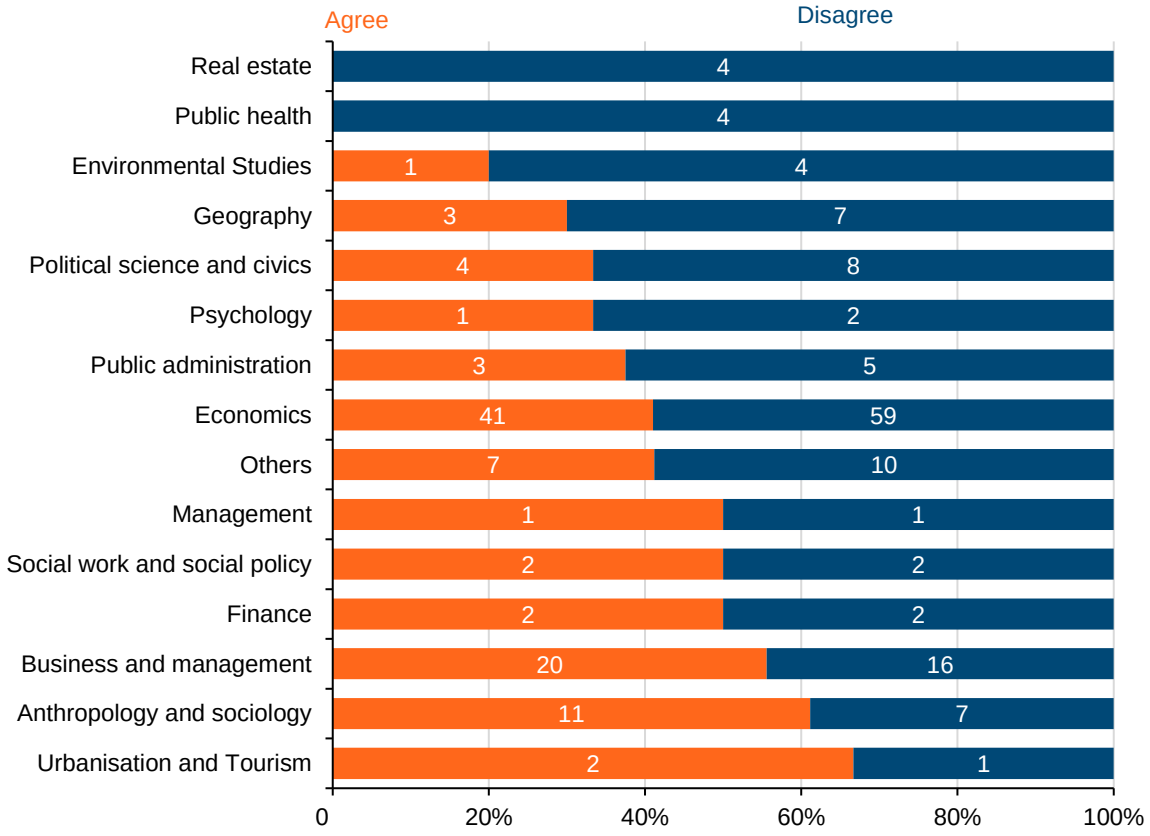
Note: Labels indicate the frequency of responses
Source: Online survey conducted by author

Figure 3.14: Response to 'In general, I think the level of open data in Malaysia is satisfactory' by years of research experience



Note: Labels indicate the frequency of responses
Source: Online survey conducted by author

Figure 3.15: Response to 'In general, I think the level of open data in Malaysia is satisfactory' based on research major



Note: Labels indicate the frequency of responses
Source: Online survey conducted by author

4. Discussion

There are two main takeaways from this survey. **First, even though data.gov.my portal is supposed to be the main platform for government data in Malaysia, it is the least popular among open data platforms.** It also has the largest percentage of respondents who regarded the platform as not useful. Most respondents would go to the relevant ministry/agency's websites to obtain data. The stated objective of data.gov.my portal, launched in 2014, is to enable easy access to open government data from one platform[14]. It seems that this is not yet the case.

Second, according to the survey, most researchers find data provided by the Malaysian government falls short in three areas: completeness, granularity and timeliness. Most respondents, however, trust the accuracy of government data and they consider them very useful for their research. In general, less than half of respondents think the level of open data in Malaysia is satisfactory.

The problem of granularity is corroborated by another survey conducted by the World Bank in September 2016 involving 232 respondents from various sectors; 89.5% of the respondents reported that data were not adequate in terms of granularity for rigorous economic research[15].

Although this survey is insightful, it has several limitations. First, since it is an online survey, the sample is small and may not be representative of the population of researchers. To mitigate this issue, I presented the responses in terms of frequency and not just in terms of percentage for transparency sake. I also avoided slicing the responses by different profile (i.e. age, institutions, years of experience, etc.) unless necessary (e.g. to investigate potential skewness) to avoid misrepresenting any particular group of researchers.

Second, this survey captures subjective perception. I do not claim that this survey is an objective evaluation of open government data in Malaysia. Several global open data evaluations, such as the Open Data Barometer and the Global Open Data Index, offer more robust methodologies for this purpose[16]. However, this survey provides insights that are not captured in global open data evaluations such as the usefulness of different open data platforms and specific issues faced by researchers in accessing government data.

Third, while this survey captures the supply side issues of open government data, it does not capture the demand side issues e.g. the data skills of researchers for meaningful use of government data. I did, however, comb out responses from respondents who are considered not familiar with open government data or who may not use government data regularly in my analysis.

[14] MAMPU (n.d.)
[15] Chuah and Loayza (2017)
[16] Refer to Ashraf (2020b)

5. Concluding remarks

This paper presents insights into the perception of open government data among researchers who conduct research on Malaysia. Less than half of the respondents rated the level of open government data in Malaysia as satisfactory. Three major areas that may need improvement are completeness, granularity and timeliness of data. The survey also highlights the low favourability of the data.gov.my portal, which is meant to be the central open government data platform in Malaysia. The portal is the last go-to open data platform according to the survey.

Cross-sectional analysis presented Box 1 in the Introduction section underscores an opportunity for countries to improve their academic research by making data open. With more data made open, researchers could conduct more meaningful analysis with greater accuracy, especially to address important issues in society. In a study conducted by The Economist, data availability of a country is one of the main predictors of the amount of research conducted on that country[17]. Therefore, for a country to get more academic research that can support policymaking, making data more available should help. Now that we are facing one of the greatest challenges of our time with the Covid-19, we need good research more than ever, not only in public health but also in other areas experiencing the far-reaching impacts of the pandemic.

[17] The Economist (2020)

6. References

Allison, Paul D., and J. Scott Long. 1990. "Departmental Effects on Scientific Productivity." *American Sociological Review* 55 (4): 469–78. https://doi.org/10.2307/2095801.

Andras, Peter, and Bruce G. Charlton. 2009. "Why Are Top Universities Losing Their Lead? An Economics Modelling-Based Approach." *Science and Public Policy* 36 (4): 317–30. https://doi.org/10.3152/030234209X436563.

Ashraf, Shaharudin. 2020a. "Covid-19: With Data, We Can Respond Better." http://www.krinstitute.org/Views-@-Covid-19-;_With_Data,_We_Can_Respond_Better.aspx.

———. 2020b. "Open Government Data: Principles, Benefits and Evaluations." Discussion Paper. Kuala Lumpur: Khazanah Research Institute. http://www.krinstitute.org/Discussion_Papers-@-Open_Government_Data-;_Principles,_Benefits_and_Evaluations.aspx.

Chuah, Lay Lian, and Norman V. Loayza. 2017. "Open Data : Differences and Implications across Countries." 114829. The World Bank. http://documents.worldbank.org/curated/en/886041494335634817/pdf/114829-BRI-Policy-7.pdf.

Conley, John P., and Ali Sina Önder. 2014. "The Research Productivity of New PhDs in Economics: The Surprisingly High Non-Success of the Successful." *Journal of Economic Perspectives* 28 (3): 205–16. https://doi.org/10.1257/jep.28.3.205.

Education First. 2020. "EF English Proficiency Index." 2020. https://www.ef.com/wwen/epi/.

Goodall, Amanda H., John M. McDowell, and Larry D. Singell. 2014. "Leadership and the Research Productivity of University Departments." Discussion Paper 7903. IZA Discussion Paper. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2385160.

Horodnic, Ioana Alexandra, and Adriana Zaiţ. 2015. "Motivation and Research Productivity in a University System Undergoing Transition." *Research Evaluation* 24 (3): 282–92. https://doi.org/10.1093/reseval/rvv010.

International Budget Partnership. 2020. "Open Budget Survey 2019." International Budget Partnership. https://www.internationalbudget.org/open-budget-survey.

International Open Data Charter. n.d. "History." *International Open Data Charter* (blog). n.d. https://opendatacharter.net/history/.

Janssen, Marijn, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. "Benefits, Adoption Barriers and Myths of Open Data and Open Government." *Information Systems Management* 29 (4): 258–68. https://doi.org/10.1080/10580530.2012.716740.

Jordan, John M., Mark Meador, and Stephen J. K. Walters. 1988. "Effects of Department Size and Organization on the Research Productivity of Academic Economists." *Economics of Education Review* 7 (2): 251–55. https://doi.org/10.1016/0272-7757(88)90049-0.

Jung, Jisun. 2012. "Faculty Research Productivity in Hong Kong across Academic Discipline." *Higher Education Studies* 2 (4): p1. https://doi.org/10.5539/hes.v2n4p1.

Kwiek, Marek. 2018. “High Research Productivity in Vertically Undifferentiated Higher Education Systems: Who Are the Top Performers?” *Scientometrics* 115 (1): 415–62. https://doi.org/10.1007/s11192-018-2644-7.

MAMPU. n.d. “MyGOV - Open Government Data - Policy, Strategy and Governance - Open Data.” The Malaysian Administrative Modernisation and Management Planning Unit. n.d. https://www.malaysia.gov.my/portal/content/30024.

Mueller, Christoph E, Hansjoerg Gaus, and Ingo Konradt. 2016. “Predicting Research Productivity in International Evaluation Journals Across Countries.” *Journal of Multidisciplinary Evaluation* 12 (27): 15.

OECD. 2015. “Making Open Science a Reality.” *OECD Science, Technology and Industry Policy Papers*, no. 25 (October). https://doi.org/10.1787/5jrs2f963zs1-en.

———. n.d. “Open Science.” Organisation for Economic Co-Operation and Development (OECD). n.d. https://www.oecd.org/science/inno/open-science.htm.

Open Data Watch. 2019. “Open Data Inventory 2018/19.” https://odin.opendatawatch.com/data/Download.

Open Knowledge Foundation. 2016. “Global Open Data Index.” Portal. Global Open Data Index. 2016. https://index.okfn.org/place/.

———. n.d. “Open Definition 2.1 - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge.” Open Definition. n.d. https://opendefinition.org/od/2.1/en/.

Rose, Carol. 1986. “The Comedy of the Commons: Custom, Commerce, and Inherently Public Property.” *The University of Chicago Law Review* 53 (3): 711. https://doi.org/10.2307/1599583.

Ryan, Matt, Dane Gambrell, and Beth Simone Noveck. 2020. “Using Collective Intelligence to Solve Public Problems.” Nesta & The GovLab. https://www.nesta.org.uk/report/using-collective-intelligence-solve-public-problems/.

Scimago Lab. 2020a. “Scimago Journal & Country Rank.” Scimago. 2020. https://www.scimagojr.com/countryrank.php.

———. 2020b. “SJR Compare Countries.” Scimago Journal & Country Rank. 2020. https://www.scimagojr.com/comparecountries.php.

Shin, Jung, and William Cummings. 2010. “Multilevel Analysis of Academic Publishing across Disciplines: Research Preference, Collaboration, and Time on Research.” *Scientometrics* 85 (2): 581–94. https://doi.org/10.1007/s11192-010-0236-2.

Sunlight Foundation. 2010. “Ten Principles For Opening Up Government Information.” Blogs. Sunlight Foundation. 2010. https://sunlightfoundation.com/policy/documents/ten-open-data-principles/.

The Economist. 2020. “Economists Look at More than GDP When Choosing Countries to Study.” *The Economist*, December 12, 2020. https://www.economist.com/graphic-detail/2020/12/12/economists-look-at-more-than-gdp-when-choosing-countries-to-study.

UNESCO. 2017. “Open Science Movement.” United Nations Educational, Scientific and Cultural Organization (UNESCO). 2017. http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/open-science-movement/.

———. n.d. “UIS Statistics.” The United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics (UIS). n.d. http://data.uis.unesco.org/.

World Bank. n.d.a. “World Development Indicators.” DataBank. n.d.a. https://databank.worldbank.org/source/world-development-indicators.

World Wide Web Foundation. 2017. “Open Data Barometer: 2016 (Fourth Edition).” https://opendatabarometer.org/assets/downloads/Open%20Data%20Barometer%20-%20Global%20Report%20-%202nd%20Edition%20-%20PRINT.pdf.

———. Various years. “Open Data Barometer Reports.” World Wide Web Foundation.

Zhang, Jinghua, Xiaoou Chen, Xin Gao, Huizeng Yang, Zhong Zhen, Qingwei Li, Yiqun Lin, and Xiyan Zhao. 2017. “Worldwide Research Productivity in the Field of Psychiatry.” *International Journal of Mental Health Systems* 11 (1): 20. https://doi.org/10.1186/s13033-017-0127-5.

Appendix A

Does open government data improve academic research output? (A preliminary study)

Academic researchers produce various types of output including patents, policy reports and media communication. However, for simplicity, this study focuses on citeable academic publications since they are the major and most easily measurable outputs of academic researchers.

The outline of this Appendix is as follows. Section A.1 reviews literature on the determinants of academic research productivity. Section A.2 describes the empirical methods employed in this study including the data used. Section A.3 provides the findings of the study. Lastly, Section A.4 discusses the findings and limitations of the study.

A.1. Literature review: Determinants of academic research productivity

Past studies have identified several determinants of academic research[18] productivity. On the individual level, a study among Romanian academics found that intrinsic motivation contributes to research productivity more than external incentives[19]. A study in Poland found that top academic researchers are likely to have strong international collaboration, produce more globally oriented research and spend longer research hours[20].

On the organisational level, a study conducted among academic economists in the United States (US) found that private institutions were more productive than public institutions[21]. Although less robust, the study also found a positive relationship between department size and research productivity. In a different study, research productivity of economics departments in the US was found to have improved when the incoming department Chair had highly cited publications[22].

A longitudinal study among academic scientists in the US showed that the reputation of a department contributes to research productivity. The reverse causal relationship (i.e. research productivity on reputation) was found to be weak[23]. On the other hand, a study on graduates from North American doctoral programmes in economics found that the rank of a department offers a poor prediction of future research productivity of graduates as opposed to their rank in their graduating class regardless of their alma mater[24].

---

[18] Recall that this study focuses on academic research and academic publication as the output. Therefore, from now on, whenever 'research' is mentioned, it refers to academic research. Also, 'research output' and 'research publication' are used interchangeably.
[19] Horodnic and Zaiț (2015)
[20] Kwiek (2018)
[21] Jordan, Meador, and Walters (1988)
[22] Goodall, McDowell, and Singell (2014)
[23] Allison and Long (1990)
[24] Conley and Önder (2014)

Some literature investigated determinants of academic research productivity at both the individual and organisational levels. A study among Hong Kong academics found that time spent on research, instruction hours for doctoral programmes, and institutions' commercial orientation contribute to research productivity[25]. Meanwhile, a study among Korean academics found that individual characteristics (e.g. academics whose primary interest lies in research and who have strong international collaborations) have stronger impacts on research productivity than institutional characteristics[26].

The literature on the determinants of research productivity at country-level is very limited, except for a few studies, which highlight the contribution of economic development to research[27]. To the best of my knowledge, no study has looked into the effect of data accessibility on research productivity at the individual, institutional or country-level. **This study, therefore, fills this gap not only with regard to country-level determinants of research productivity but, more importantly, the role of open government data.**

A.2. Methods

This study uses economic tools in modelling research production to systematically establish the role of open government data.

Empirical estimation model

I introduce a simple cross-country research production model. Suppose there are two factors[28] in research production, namely scientific labour (i.e. researcher) and capital investment (i.e. research expenditure excluding labour costs). Consider these two factors as non-substitutable, which means an increase in one factor without a corresponding increase in the other factor will not increase the output (i.e. the academic publication). For example, purchasing more microscopes without increasing the number of researchers is not likely to improve significantly the level of academic publications, nor vice versa.

Non-substitutability, however, does not mean that the two factors are needed in the same amount to produce a certain level of output. For example, a higher number of researchers than the number of microscopes may be needed to produce a certain level of academic publication but neither factor can be substituted by the other.

The approach taken in this study is different than past studies, which neglect the non-substitutability of labour and capital in research production[29]. Why does this matter then? Unlike activities such as farming and manufacturing in which hard labour can be easily replaced with a machine, there is no satisfactory substitute for human intelligence in research production, at least for now. Of course, the invention of many technologies, such as computers, the internet,

[25] Jung (2012)
[26] Shin and Cummings (2010)
[27] Mueller, Gaus, and Konradt (2016) & Zhang et al. (2017)
[28] In economic terms, 'factors' can be loosely defined as 'input'. Think of them as the necessary items to produce something.
[29] Andras and Charlton (2009) & Zhang et al. (2017)

microscopes and satellites, have improved research over the years. However, these technologies enhance research process and outcome instead of substituting human intelligence, the main contribution of scientific labour.

Likewise, some technologies cannot be substituted with scientific labour. For example, even the best scientific labour could not replace the function of a microscope. Hence, both labour and capital are needed in research production for their non-substitutable functions. My model embodies this conceptual clarity, which previous studies did not.

The research production function described above can be mathematically written as equation (1) and graphically depicted as Figure A.1. Capital investment (i.e. capital expenditure excluding labour costs) is denoted by K whereas scientific labour (i.e. researcher) is denoted by L. The factor multipliers of K and L are represented by α and β, respectively. The level of output (i.e. academic publication) is denoted by Q.

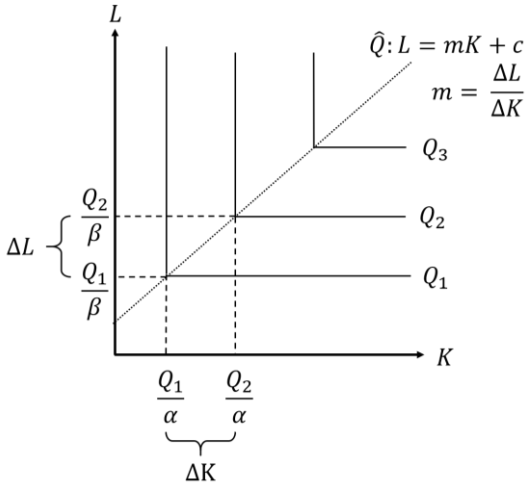
Refer to Figure A.1. Say we have Q_1/α of K and Q_1/β of L. This results in the level of output of Q_1. An increase in K up to Q_2/α without an increase of L to at least Q_2/β will not shift the level of output upward to Q_2; the level of output will remain at Q_1.

In equation (1), γ is the rate of return of L and K, which refers to the rate of increment of Q (e.g. from Q_1 to Q_2) with the increment in the combination of αK and βL. Besides γ, the rate of which Q_1 shifts to Q_2 also depends on several variables, called drivers in this model, which are represented by the matrix p with the rate of return μ.

(1) $$Q = f\,(K, L, p) = [\min\,(\alpha K + \beta L)]^{\gamma} p^{\mu}$$

where Q: Research output; K: Research capital; L: Researchers; p: Drivers of research production; γ and μ: Rate of returns

Figure A.1: Research production function



Source: Author's illustration

As a cross-country model, the different levels of Q are represented by the output of different countries. Assume that countries optimise the allocation of K and L in research production, which means countries will not excessively allocate K without a corresponding level of L and vice versa.

Hence, the combinations of K and L of different countries will hover around the line $\hat{Q}$ in Figure 2.1, whose relationship can be written as equation (2).

(2) $\hat{Q}: L = mK + c$
where c: Intercept

From Figure A.1, it can be shown that the slope of $\hat{Q}$, denoted by m, is the ratio of α to β. Following the assumption that countries optimise their factors of research production, the minimum function can be removed from the equation (1). Next, substituting m and c into the equation (1) gives equation (3). Taking the log of both sides of equation (3) gives equation (4).

(3) $$Q = (\alpha K + \beta L)^{\gamma} p^{\mu} = \left[\left(m + \frac{c}{K}\right)\beta K + \beta L\right]^{\gamma} p^{\mu} = \beta^{\gamma}(mK + c + L)^{\gamma} p^{\mu}$$

(4) $$\ln(Q) = \tau + \gamma \ln(mK + c + L) + \mu \ln(p)$$
where $\tau = \gamma \ln \beta$

Two drivers of research production are considered in this study: the country's economic development and open data[30] level. The selection of economic development as another driver of concern in this study, apart from the open data level, is because it may capture a lot of cross-country variations including the quality of researchers and pre-existing research infrastructure.

The final estimation model is written as equation (5).

(5) $$\ln(Q) = \tau + \gamma \ln(mK + c + L) + \mu_1 \ln(dev) + \mu_2 \ln(od)$$

Data sample

As previously mentioned, research output in this study refers to academic publication. Although academic publishing is mainly carried out by the higher education sector, I also conduct the analysis for the aggregate of all sectors (including business, non-profit, and public sectors). This is because while the number of researchers in the higher education sector in most countries is more than half of the total number of researchers of all sectors, the research expenditure (excluding labour costs) from the higher education sector accounts for less than half of the total research expenditure (Table A.1).

Table A.1: No. of countries based on the share of the higher education research expenditure (excluding labour costs) and number of researchers out of the total of all sectors (sample countries = 76)

Share of research expenditure (no. of researchers) of the higher education sector over the total research expenditure (no. of researchers) of all sectors	No. of countries (share of the total sample)	
	Research expenditure	No. of researchers
< 25%	29 (38%)	5 (7%)
25 – 50%	36 (47%)	20 (26%)
50 – 75%	9 (12%)	32 (42%)
> 75%	2 (3%)	19 (25%)
Total sample	76 (100%)	76 (100%)

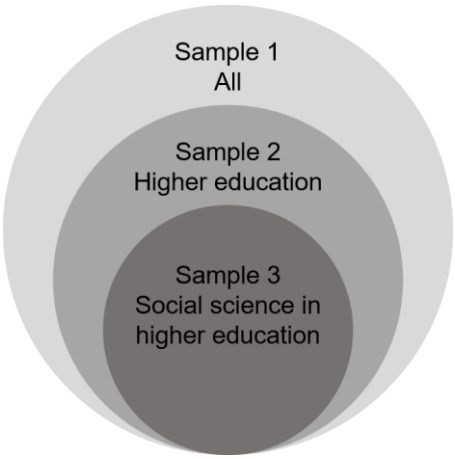
Source: UNESCO (n.d.)

[30] It is worth reiterating that the focus of this paper is on open government data. Therefore, open data and open government data are used interchangeably.

The use of government data may be more prevalent in social science research as opposed to natural science research. This is because the latter may depend more on primary data such as laboratory and field data. While I do not have conclusive evidence for this assumption, I decide to err on the side of caution and account for this potential bias by also conducting the analysis on social science research in the higher education sector.

In short, analysis in this study is conducted for three samples: research in all sectors (sample 1), research in the higher education sector (sample 2) and social science research in the higher education sector (sample 3). Sample 3 is a subset of sample 2, which is a subset of sample 1 (Figure A.2).

Figure A.2: Data samples in the study



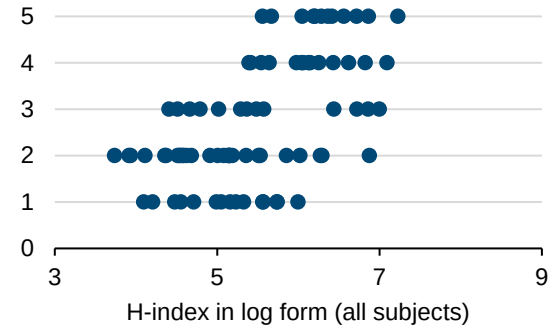
Source: Author's illustration

Variables and data source

Research output (*Q*) is measured by h-index, which is each country's number of academic articles (*h*) that have received at least *h* citations, based on Scopus database. Scopus is one of the largest academic publications database covering 11,678 publishers. H-index represents not only the number of research publications but also their quality. The average index between the years 2013 and 2016 is taken in this study. The selection of years up to 2016 is because the latest Open Data Barometer score, which is used for the open data variable in this study and will be elaborated further, is only available up to 2016. *Q* data of sample 1 and 2 that covers all subjects is different than sample 3 that only covers social science subjects.

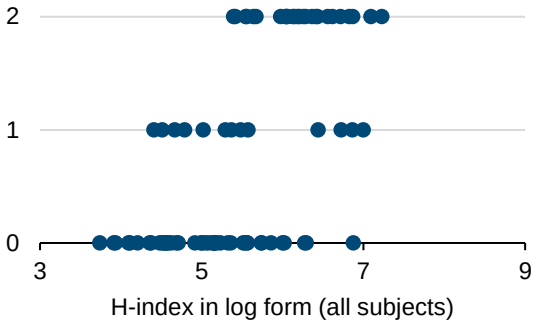
To account for the Scopus bias towards English language publications, a dummy variable of English language proficiency (*eng*) is included in the estimation. Countries are divided into three groups based on 2016 data from the Education First's English Proficiency Index (EPI): (i) countries that have very high and high proficiency, (ii) countries that have moderate proficiency, and (iii) countries that have low and very low proficiency. For countries of which data is unavailable, data for other years is taken, or self-imputation is made (e.g. countries with English as the native language are assigned very high proficiency). The five levels of English proficiency categorised by the Education First data are regrouped into three since there seem to be little difference in h-index of countries with very high and high proficiency level as well as countries with low and very low proficiency level (Figure A.3 & A.4).

Figure A.3: Countries' h-index (for all subjects) versus their English proficiency level based on Education First's five-level categorisation



Note: Proficiency level: 1: Very low to 5: Very high. H-index is plotted on the x-axis for better visualisation.
Source: Education First (2020)

Figure A.4: Countries' h-index (for all subjects) versus their English proficiency level based on three-group recalibration



Note: Proficiency level: 0: Very low and low to 2: High and very high. H-index is plotted on the x-axis for better visualisation.

Data for the number of researchers (L) and research expenditure excluding labour costs (K) are sourced from the United Nations Educational, Scientific and Cultural Organisation (UNESCO) Institute for Statistics (UIS). L and K are normalised via log transformation for better comparability between the two variables.

To calculate K, labour costs are subtracted from the total research expenditure to prevent endogeneity issues with L since the larger the number of researchers, the higher the expected labour costs. Since labour costs data is not available for sample 3 (i.e. the social science in the higher education sector), it is assumed that each country's share of labour costs out of the total research expenditure in this sample follows sample 2 (i.e. the higher education sector).

For countries whose labour costs data is not available at all, it is assumed that 54% (in sample 1 & 3) and 57% (in sample 2) of their total research expenditure comprise of labour costs. This follows each sample's median share of labour costs out of the total research expenditure based on countries with available data. The number of countries which labour costs are not available is shown in Table A.2.

Table A.2: Countries with no data on labour costs

Sample	Total countries	No. of countries with no data on labour costs	Share of countries with no data on labour costs
Sample 1	83	27	32.5%
Sample 2	76	20	26.3%
Sample 3	51	*With no data from sample 2 as a proxy: 8	15.7%

*Note: Labour costs are not available for sample 3 (i.e. social science in the higher education sector) for all countries. Except for 8 countries, the share of labour costs of the remaining countries in sample 3 follows their respective share in sample 2.

The values of L and K are different for sample 1, 2 and 3. Figure A.5 – A.7 show the relationship of L, K and Q (all expressed in log form). There are two things to pay attention to in these figures. First, L and K exhibit a linear relationship and Q generally increases with the increase in L and K. This indicates the non-substitutability of L and K in producing Q. The counter scenario is an inverse relationship between L and K (that indicates substitutability of the two factors) or no relationship exhibited at all. Second, notice that some countries have a higher level of Q compared to other countries that have the same combination of L and K as them. This means that L and K are not the only determinants of Q; there are other determinants, which I call drivers in my model, denoted by p (refer equation (4)).

Figure A.5: Relationship of K, L and Q in sample 1

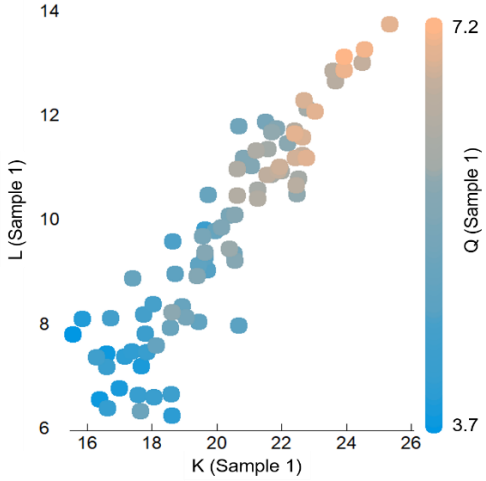


Figure A.6: Relationship of K, L and Q in sample 2

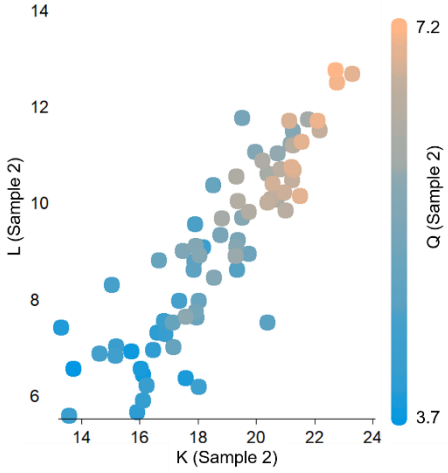
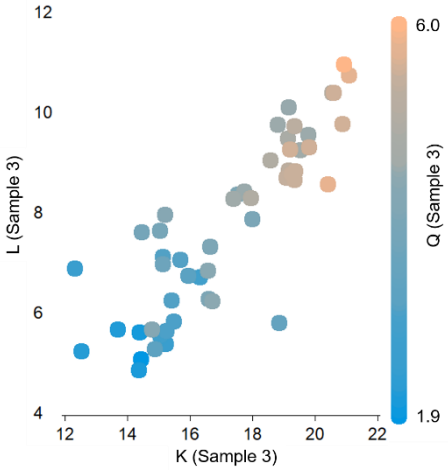


Figure A.7: Relationship of K, L and Q in sample 3



The level of economic development (*dev*) is represented by the average GDP per capita whereas the level of open data (*od*) is represented by the Open Data Barometer (ODB) score. The ODB, carried out by the World Wide Web Foundation, measures the prevalence and impact of open data provided by the public sector. The ODB uses an in-depth methodology[31] that combines contextual data, technical assessments and secondary indicators as follows:

- **Peer-reviewed expert survey:** Carried out with a range of questions about open data contexts, policy, implementation and impacts and a detailed dataset survey completed for 15 kinds of data, which touch on issues of data availability, format, licensing, timeliness and discoverability.
- **A government self-assessment simplified survey:** With the same range of context, implementation, and impacts questions to supplement the expert survey.
- **Secondary data:** Used in the readiness section of the Barometer and taken from the World Economic Forum, International Telecommunications Union, United Nations e-Government Survey, and Freedom House.

[31] World Wide Web Foundation (Various years)

Table 2.3: Variables description and sources

Variable	Indicator	Description	Source
Q	H-Index	Average h-index (for all subjects and social science) between 2013 and 2016	Scopus database, Scimago Lab (2020)
K	Research expenditure excluding labour costs	Average research expenditure excluding labour costs (for all sectors, higher edu. sector, and social science in higher edu.) between 2013 and 2016 (log-transformed)	UNESCO (n.d.)
L	Researchers	Average no. of researchers (for all sectors, higher edu. sector, and social science in higher edu.) between 2013 and 2016 (log-transformed)	UNESCO (n.d.)
dev	GDP per capita	Average GDP per capita between 2013 and 2016	World Development Indicators, World Bank (n.d.a)
od	Open Data Barometer (ODB) score	Average scaled score between 2013 and 2016	World Wide Web Foundation (Various years)
eng	English Proficiency Index	2: Native, very high proficiency, high proficiency; 1: Moderate proficiency; 0: Low proficiency, very low proficiency	Education First (n.d.). Self-imputed for countries that do not have data.

Note: Average figures between 2013 and 2016 are taken for all variables except *eng* to mitigate the issue of data gaps.

Preliminary analysis

The skewness of all variables except $\ln(od)$ are between -0.5 and 0.5, which means that they are approximately normally distributed (Table 2.4). The skewness of $\ln(od)$ is less than -1, which means it is moderately skewed. Sample countries are chosen based on data availability. The number of observations decreases with the increase in granularity, i.e. N (sample 1) > N (sample 2) > N (sample 3), due to less data being available.

Table 2.4: Descriptive statistics of variables

Variable	Mean	Median	Std. Dev.	Skewness
Sample 1 (N = 83): All sectors				
K	20.19	20.40	2.33	-0.02
L	9.72	9.80	1.98	-0.04
$\boldsymbol{\ln(Q)}$	5.48	5.48	0.86	0.05
$\boldsymbol{\ln(mK + c + L)}$*	2.95	2.96	0.20	-0.21
$\boldsymbol{\ln(dev)}$	9.02	9.26	1.42	-0.39
$\boldsymbol{\ln(od)}$	3.29	3.33	0.72	-0.50
Sample 2 (N = 76): Higher education sector				
K	18.79	19.06	2.35	-0.31
L	9.11	9.10	1.87	-0.06
$\boldsymbol{\ln(Q)}$	5.48	5.50	0.89	0.06
$\boldsymbol{\ln(mK + c + L)}$*	2.88	2.91	0.19	-0.40
$\boldsymbol{\ln(dev)}$	9.08	9.23	1.45	-0.43
$\boldsymbol{\ln(od)}$	3.32	3.35	0.72	-0.63
Sample 3 (N = 51): Social science higher education sector				
K	17.24	17.39	5.62	-0.11
L	7.72	7.85	1.72	0.02
$\boldsymbol{\ln(Q)}$	4.08	4.09	0.98	-0.19
$\boldsymbol{\ln(mK + c + L)}$*	2.71	2.68	0.04	-0.17
$\boldsymbol{\ln(dev)}$	9.31	9.55	1.39	-0.57
$\boldsymbol{\ln(od)}$	3.41	3.53	0.65	-0.38

Note: *Descriptive statistics for $\ln(mK + c + L)$ is obtained after identifying the value of *m* and *c* as described in Section A.3.

A.3. Results

Table A.5 presents the regression results of L on K, which give the values of m and c. These values are substituted into the term $\ln(mK + c + L)$ in equation (5). This term essentially represents the compounded two-factor variable in research production.

Table A.5: L on K estimation results

	Sample 1	Sample 2	Sample 3
No. of observations	83	76	51
Coefficients:			
m	0.7818*** (0.0369)	0.6996*** (0.0468)	0.6256*** (0.0602)
c	-6.0686*** (0.7890)	-4.0379*** (0.9155)	-3.0669*** (1.0765)
R^2	0.8473	0.8918	0.7395

Note: Asterisks *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Robust standard errors are in parentheses.

Table A.6 presents the estimation results of equation (5). To mitigate the heteroskedasticity issue, robust standard errors are reported. The Shapiro-Wilks normality test ran found that the residuals for sample 2 are not normally distributed. Nevertheless, there is no multicollinearity issue among variables, as verified through the variance inflation factor (VIF) test.

Table 2.6: Estimated effects of research production factors and drivers on research output

	Sample 1		Sample 2		Sample 3	
	(A)	(B)	(C)	(D)	(E)	(F)
Coefficients of:						
$\ln(mK + c + L)$	2.8503*** (0.2113)	2.7862*** (0.2039)	2.8393*** (0.2246)	2.7410*** (0.2247)	3.4795*** (0.4318)	3.4629*** (0.4238)
$\ln(dev)$	0.1254*** (0.0370)	0.1070*** (0.0374)	0.1106** (0.0524)	0.0825* (0.0480)	0.0413 (0.0745)	0.0150 (0.0717)
$\ln(od)$	**0.2377*** (0.0775)**	**0.2024*** (0.0735)**	**0.2666** (0.1351)**	**0.2238* (0.1311)**	**0.2937** (0.1391)**	**0.2050 (0.1498)**
eng	-	0.1011*** (0.0373)	-	0.1429*** (0.0539)	-	0.1390** (0.0686)
Intercept (τ)	-4.8339*** (0.5072)	-4.4420*** (0.4933)	-4.5994*** (0.6277)	-4.0354*** (0.6833)	-6.7560*** (0.9334)	-6.3077*** (0.8948)
Sample size	83	83	76	76	51	51
R^2	0.8933	0.9001	0.8552	0.8680	0.8577	0.8659

Note: Asterisks *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Robust standard errors are in parentheses.

Open data is found to be statistically significant with a positive coefficient for all three samples except with the inclusion of *eng* variable in sample 3. However, economic development is also found to be statistically insignificant in sample 3. This may be because the size of sample 3 is small and for countries in the sample, research output is almost entirely accounted for by the factors of production, namely scientific labour and capital investment.

A.4. Discussion

The regression results show that, controlling for factors of research production and the level of economic development, the higher the level of open data, the higher the h-index of academic publications. This finding is consistent across three samples: research in all sectors, research in the higher education sector, and social science research in the higher education sector.

Undoubtedly, researchers and capital investment are the factors of research, hence they are the main determinants of research output. Rather interestingly, an improvement in the GDP per capita and the ODB score would render a comparable improvement in the h-index. This highlights the significance of open government data for academic research.

This study, of course, suffers several limitations. First, since the study employs a cross-sectional analysis, it does not capture the time dynamic aspect. If country X improves its open data level, would country X be able to improve its academic output over time? What I show in this study is country Y that has a higher level of open data has better academic output than country X that has a lower level of open data. A panel data analysis could show whether improvement in open data drives research production over time. However, this is not carried out in this study for two reasons: (i) only several countries recorded tremendous improvement in their ODB scores over the years and (ii) there are only four ODB instalments so far (2013 – 2016) while research improvement may require a longer time.

Second, it is difficult to obtain a completely objective measurement of open data. This is because all established open data evaluations, including ODB, rely on an expert survey. With the robust methodology that ODB employs, I am confident that ODB score is the best measurement of open data level that I could use for now.

Third, given the borderless nature of research production whereby researchers from different countries can collaborate to produce a research output or countries can share costs to fund a research project, trying to capture scientific labour and capital investment within the confines of a country may suffer measurement issue. Due to the limitations of data, I was not able to account for this.

Appendix B

Survey questionnaire

Perception of academics towards open data in Malaysia

Page 1

Khazanah Research Institute (KRI) is conducting a study on open data in Malaysia. The objective of this survey is to understand the perception of open data among researchers, based locally or abroad, who do research in Malaysia.

This survey is especially relevant for social science researchers. However, response from researchers in other fields who use open data, such as public health and environmental studies, is highly welcome.

Open data is defined as "data that can be freely used, shared and built-on by anyone, anywhere, for any purpose" – Open Knowledge Foundation.

The survey consists of 6 pages of short multiple-choice questions and should take you approximately 5 - 7 minutes to answer.

The summarised outcome of this survey will be published and used for policy recommendations. However, any personally identifiable information will remain confidential.

Your participation is voluntary, and you can withdraw from answering this survey at any point.

If you have any questions regarding this survey, please feel free to contact Ashraf Shaharudin via email: ashraf.shaharudin@KRInstitute.org

Thank you for participating in this survey.

Abbreviation:
DOSM: Department of Statistics Malaysia

Page 2

What is your current role? (Choose the most relevant)

- Academic staff
- Postgraduate student
- Researcher in a higher education institution but neither of the above
- Researcher in a non-higher education institution

Which type of institution are you mainly affiliated with?

- Public higher education institution (Local)
- Private higher education institution (Local)
- Non-higher education institution (Local)
- Higher education institution not based in Malaysia
- Non-higher education not based in Malaysia

What is your research major? (Choose the most relevant one)

- Anthropology and sociology
- Political science and civics
- Economics
- Business and management
- Public administration
- Geography
- Public health
- Environmental Studies
- Others

What is your age range?

- < 18
- 18 – 25
- 26 – 35
- 36 – 45
- 46 – 55
- > 55

How long have you been a researcher?

- < 1 year
- 1 – 3 years
- 4 – 6 years
- > 6 years

Page 3

Have you heard of these data platforms?
'Yes' or 'No'

- Data.gov.my
- eStatistik, DOSM
- World Bank Open Data
- IMF Data
- Un Statistics Division

How useful are these platforms for you in obtaining data? (This question has 4 options. If you are using a mobile phone, scroll sideways to see all options.)

Rank 1 – 4, 1 being the most useful

- Data.gov.my
- eStatistik, DOSM
- World Bank Open Data
- IMF Data
- UN Statistics Division

Page 4

Which platform would you go first to obtain these data? (1/2) (This question has 6 options. If you are using a mobile phone, scroll sideways to see all options.)

1 = Data.gov.my, 2 = DOSM website, 3 = Ministry/Agency's website (other than the first two), 4 = International organization database, 5 = Not sure, 6 = Not relevant for me

- Map data (eg. borders, roads, etc.)
- Land ownership
- Economic indicators
- Federal government budget
- Federal government expenditure
- Company registration
- Legislation data

Page 5

Which platform would you go first to obtain these data? (2/2) (This question has 6 options. If you are using a mobile phone, scroll sideways to see all options.)

1 = Data.gov.my, 2 = DOSM website, 3 = Ministry/Agency's website (other than the first two), 4 = International organization database, 5 = Not sure, 6 = Not relevant for me

- Public transport information
- International trade data
- National health & morbidity data
- Education data
- Crime statistics
- National environment statistics
- National election results
- Public procurement

Page 6

What is your perception of **open data provided by the Malaysian government** through platforms such as data.gov.my and agency/ministry websites based on these criteria? (1/2)

1 = Agree, 2 = Disagree, 3 = Not sure

- Most data that I want is readily available online.
- I trust the accuracy of most data obtained through government platforms
- Most data provided online comes with complete metadata that helps explain the data
- I rarely need to request additional information on the data from the relevant agency/ministry
- Most online data will not be taken down after some period
- I can obtain most old data such as from the 1980s online
- Most data provided online is down to the level of granularity that I want
- Most data provided online is up to date

Page 7

What is your perception of **open data provided by the Malaysian government** through platforms such as data.gov.my and agency/ministry websites based on these criteria? (2/2)

1 = Agree, 2 = Disagree, 3 = Not sure

- Most data is made open free of charge
- Most data is easily downloadable in bulk
- Most data is provided under an open license
- Most data is provided in a machine-processable format, such as Excel or CSV and not PDF
- I do not have to purchase or download specific software to obtain most data
- I do not have to register to obtain most data
- Most data that I obtain from several different sources are easily comparable
- Most data provided online is very useful for me to do my research
- In general, I think the level of open data in Malaysia is satisfactory