Measuring citizen preferences for the Dutch Education Open Data Policy: A Path towards Citizen-Informed Decision Making

Darli Ciang

Master Thesis

August 2018

Measuring citizen preferences for the Dutch Education Open Data Policy: A Path towards Citizen-Informed Decision Making

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Complex Systems Engineering and Management

Faculty of Technology, Policy and Management

by

Darli Ciang

Student number: 4624211

To be defended in public on August 27th, 2018

Graduation committee

Chairperson	: Prof.dr.ir. M.F.W.H.A. (Marijn) Janssen, Section ICT
First Supervisor	: Prof.dr.ir. M.F.W.H.A. (Marijn) Janssen, Section ICT
Second Supervisor	: Dr. Mr. N. (Niek) Mouter, Section T&L
Third Supervisor	: Dr. A.M.G. (Anneke) Zuiderwijk- van Eijk, Section ICT

Cover photo: Code for NL launch
"Workshop: code for method" by Code For NL launch, Waag Society is licensed under CC BY-NC-SA 2.0

Preface

This master thesis concludes my master program Complex Systems Engineering and Management at the Technology, Policy and Management faculty of the Delft University of Technology. The master thesis graduation project is aimed to empirically measured citizens preferences of Dutch open education data attributes using the citizen stated choice experiment. In this special occasion, I would like to express my deepest gratitude to those who have made this thesis possible.

First, I would like to truly thank my graduation committee for guiding and supporting me in the journey to conclude my master study. Anneke Zuiderwijk, who supervises and challenge me to keep improving this graduation project. I am grateful for our bi-weekly meeting, thank you for asking about my condition. It means a lot when I faced a slump in my progress. Niek Mouter, who introduces me to the citizen stated choice experiment and always available for discussion. Thank you for the insight and your patience to guide me who had zero experience in conducting stated preference study. Marijn Janssen, who is willing to become the chairman and review my progress on top of his busy schedule. Thank you for your valuable feedback and your flexibility that you agree to check my report during your vacation.

To my Indonesian friends in daily life as students: Samuel, Rina, Bramka, Rosa, Ayu, and Timothy. Thank you for your companionship for these two years. I will treasure the memories that we made: Wintervakantie, game and movie nights, and even the study night before the exam.

To my fellow CoSEM colleagues: Lars, Inez, Daan, Mike, and Sam. Our casual meeting in TPM and the discussion that we have, recharge me for the needed social interaction. Having all of you as friends motivate me to finish this master study.

To my friends from International Christian Fellowship (ICF) Delft: Yosua, Ina, Arjan, Aliya, Vitali, Febe, Vincent, and Nikola. Thank you for your prayer and fellowship during my stay in Delft.

I also want to give my gratitude to LPDP (Indonesia Endowment Fund for Education) for the scholarship that enables me to pursue this master study. Thank you for giving me a chance to broaden my knowledge.

Lastly, to people who are dearest to me. My parents, *Papa* and *Mama*. Seeing your face and calling you at the weekend is one of my favorite moment in this journey. Finally, after two years I can finish my study and meet both of you in Indonesia.

Darli Ciang
Delft
August 2018

Executive Summary

This report discusses the way individuals in their role as citizens make trade-offs between open education data attributes. The study is conducted to address the problem of “lack of insight in the citizen preferences of open data policy attributes”. This lack of insight has influenced policymakers on how they develop and evaluate open education data policy. In the current situation, government agencies tend to replicate “best practice” policy from other agency without considering their policy objectives and context. This tendency to mimic other agency and lack of insight on the citizens preferences lead them to evaluate and develop their open data policy only from the data provider perspective. In order to address the problem, this research aims to identify the citizens preferences for the open data policy through the citizens stated choice experiment (CSCE).

Two dominant valuation methods are revealed preference and stated preference method. However, in many public goods such as environmental valuation, human health effects, and other outcomes for which (direct or indirect) revealed preference (RP) data are not available; stated preference methods are the only known approach to estimate values for changes. In this study, we select a variant of stated preference method called citizen stated choice experiment (CSCE). CSCE is preferred over the consumer stated choice experiment due to the reality of open data policy implementation which is fully-funded from the government budget and provided without any charge for its utilization. The Dutch government also stated that open public data can be re-used without restriction in the form of cost, compulsory registration. Therefore, there is no scenario for consumer preferences in the current open data policy context. The citizen stated choice experiment is designed in the form of a survey with narratives and choice tasks. Five steps of conducting the discrete choice experiment are explained: 1) establishing attributes, 2) assigning attribute levels, 3) designing the choice sets, 4) generating, pre-testing, and distribute the questionnaire, and 5) analyze DCE data.

Next, the literature review of open data policy study is conducted to identify the possible trade-off attributes for the survey. Three potential categories of attributes are identified from the literature review: data-related attributes, portal-related attributes, and participation & engagement attributes.

These three attributes are further refined using information collected from the policy context of open education data (policy objectives, organization context, and existing implementation). Data-related attributes and participation & engagement attributes are identified in the policy context. In the policy context exploration, one particular aspect of data-related attributes is considered very important in the policy context which is the data protection attribute. Therefore, the three categories of open education policy attributes are modified into data-related attributes, data protection attribute, and participation & engagement related attributes as shown in Figure 1. The portal-related attributes are omitted because in the context of open education data policy there is only a basic open data portal implementation.

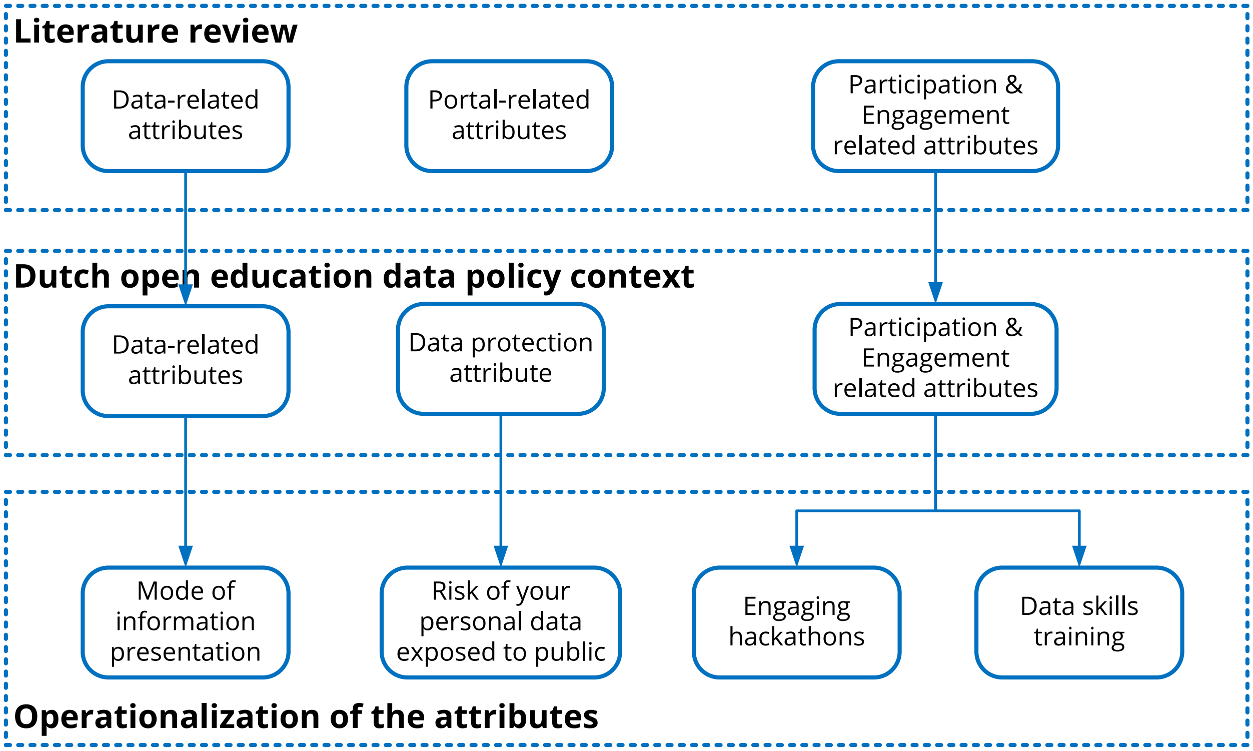

Literature review
Data-related attributes
Portal-related attributes
Participation & Engagement related attributes
Dutch open education data policy context
Data-related attributes
Data protection attribute
Participation & Engagement related attributes
Mode of information presentation
Risk of your personal data exposed to public
Engaging hackathons
Data skills training
Operationalization of the attributes

Figure 1 Conceptual framework of open education data policy attributes

We based the selection of the attributes on three criteria: the expected influence on an individual, the societal relevance of the factor, and measurability in the discrete choice experiment. The orthogonal design is favored over D-efficient design due to its robustness for an experiment without established prior values (estimated parameters found from the previous study). The D-efficient design is the most efficient for an experiment if the prior values used to generate the design is accurate and close to the true values. However, it is risky to use D-efficient design with uncertain prior because the design significantly becomes inefficient when the true values deviate from the prior values. The analysis of the citizen stated choice experiment is conducted based on the Lancaster's characteristic demand theory and Random Utility Theory. Combining both theories enables the researcher to explicitly estimate citizen preference of attributes (β) based on the alternative chosen by a citizen in the designed choice situations.

Next, the attributes were quantified by determining their unit of measurement and available attribute level ranges. The attributes and attribute levels become the input for the survey design. Table 1 highlights the attributes, attribute levels and ranges of all attributes.

Table 1 Overview of attributes and attribute levels

Category	Attributes	Value
Data-related attribute	Mode of information presentation	• in original form (as similar as possible to the source) • as static or interactive figures

		• as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)
Participation & engagement related attribute	Number of free engaging hackathon events	• 1 every 2-years • 1 per year • 2 per year
	Number of free citizen data skill training events	• 1 per year • 2 per year • 3 per year
Data protection attribute	risk of your personal education data exposed to the public	• 1 incident per year • 1 incident every 3-months • 1 incident per month

The final survey is distributed among students who are currently attending Dutch higher education institutions or recently graduated. The higher education students are targeted due to several reasons:

1. they have relevant use case for the open education data which make them more likely to know about open data,
2. they have relevant skills to use open education data which make them more likely to be motivated on using open education data
3. they are more likely to understand the term used in the survey with a proper explanation compared to other potential respondents (i.e., parents, primary/secondary school students).

In total 531 observations are collected from 59 respondents who completed the online survey (each respondent complete 9 choice situations). Majority of the respondents are familiar with open education data portals and services created using open education data. 64% of the respondents have visited at least 1 open education data portal in their life and 61% of the respondents has used at least 1 service created using open education data, such as studiekeuze123.nl and scholenopdekaart.nl. However, only 7% of the respondents has attended open education data events (e.g., “Hack de Valse Start” and “Onderwijsdata onder de loep”) in their life.

A Multinomial Logit (MNL) model is used to analyse citizens preferences for the main attributes. Citizens derive significant utility for “*mode of information presentation*” and “*risk of your personal education data exposed to the public*” attributes. The *“mode of information presentation”* is a non-linear attribute, significant improvement of the utility is shown when the information is presented as a service compared to the information as a figure and information in the original form.

The citizens significantly valued open education data policy with lower “*risk of your personal education data exposed to the public*” and the impact of higher data leak incident rate can offset the utility gain from the improvement in other open data policy attributes and dominate their choices. It explains the existence of dominant choice in the descriptive result of choice distribution.

However, the possibility of 'hypothetical bias' should be considered in the interpretation of the result. In the survey, open education data breach is described as follow: "The open education data is anonymized. The personal data leakage happens when a person can be identified from the combination of multiple anonymous open datasets". 58% of the respondents shows no concern about the possibility of data privacy breach. It seems in reality respondents have less concern on the possibility of data breach and the wording of choice situation exaggerate the chance. The policymaker should consider this fact in the interpretation of this study and further investigation is needed to conclusively determine the utility of data protection attribute.

Other than that, two attributes are considered insignificant by the citizens *"number of free citizen data skill training events"* and *"number of free engaging hackathon events".* However, the descriptive result shows that only 7% of the respondents have attended open education data events. It might have been difficult for respondents to assess their preferences for participation and engagement events (hackathon/data skills training) if they have never attended one.

Given the citizens reluctance to compromise the data protection attribute, government agencies have limited option for the implementation. Two recommendations are formulated to improve the existing open education data policy:

1. Collaborate with infomediaries (users who create services from open data for end users) to provide services for citizens
2. Engage the citizens in a cost-efficient manner.

We recommend future research to collect primary source information (interview) from the policymakers in order to improve the realism and obtain detail information that are not published in the publicly available policy documents. Other than that, we suggest including the cost for the implementation of each attribute in the survey design. It will be interesting to investigate whether respondents valuate the trade-off attributes differently if the cost of the implementation is revealed.

In this research, the experiment is limited to open education data and higher education students as the target respondents. Future research can explore different policy context (e.g., open data policy for geospatial data, science data) or different respondents for open education data. For example, open education data policy for primary and secondary schools which targets the parents and pupils as the users.

Finally, in reality there are many attributes that can be included or combined to make different alternatives. The portal-related attribute is omitted from this study because the limited implementation of open data portal in the Dutch open education data. However, if the future research explores the portal-related attributes of city open data portal, the attributes selected will be different from the attributes in this study. The attributes can focus on the functionality and features of the open data portal such as the visualization capability, collaboration and communication features, format of the data provided compared to the socio-technical perspective of this study.

Table of Contents

List of figures

List of tables

Chapter 1: Introduction

In recent years, governments throughout the world have adopted open data policies. Several countries spearheaded the initiatives such as the United States with Open Government directive and Digital Government Strategy under Obama administration (Obama, 2009, 2012) and the European Union with Directive 2013/37/EU about the reuse of public sector information and the European Commission's Open Data Strategy (European Commission, 2003, 2011). These policies aim for transparency, participation, citizen-government collaboration, evidence-based policy making, administrative efficiency, stimulate innovation, and economic growth (European Commission, 2011; Obama, 2009).

1.1. The context of Dutch open education data policy

The vision of Open Government in the Netherlands

The Open Data initiatives in the Netherlands started in parallel with the Dutch government greater commitment for Open Government through the publication of *Strategic Vision on Open Government* and *Open Government Action Plan* in 2013 (Rijksoverheid, 2013b, 2013a). In 2014, the Parliament amended the Public Access Act (WOB) which changes the stances of government from passively responds to the citizen data requests into actively open up the government data to the public (Algemene Rekenkamer, 2014). Furthermore, the Reuse of Public Information Act (WHO) provides the legal framework for the reuse of public data (data.overheid.nl, 2018).

The Open Government Action Plan contains a series of actions and the respective government agencies that are responsible for each action to achieve the Open Government vision. It is updated every two years, and the third Open Government Action Plan 2018-2020 is currently being prepared.

The open government action plan focuses on several key points (Rijksoverheid, 2018a):

1. Improving access to government information.
2. Public accountability
3. Promote openness in the government and active collaboration with the public.

One of the action points in Open Government Action Plan is the creation of a National Open Data Agenda (NODA) with the goal to increase the availability of open datasets on the national open data portal data.overheid.nl (Ministry of the Interior and Kingdom Relations, 2017). In order to achieve this goal, the government-wide inventory of datasets is conducted; Ministry of the Interior and Kingdom Relations (BZK) is given the tasks to monitor the progress of publishing open datasets and assisted other government agencies in the process of publishing open datasets.

The Ministry of the Interior and Kingdom Relations (BZK) commissioned The Open Government Learning and Expertise Point (LEOO) as the knowledge broker for the advancement of the action plan. It acts as a facilitator for public professionals who want to know, engage, and participate in one or more of the Open Government aspects: Open Contact, Open Approach, Open Data and Open Accountability (open-overheid.nl, 2018).

Open Education Data

The Open Public Data is defined as data that: are paid for from the public purse and generated during or for the provision of a public service, are available to the public, are free of copyright and other third-party rights, are machine-readable and preferably comply with open standards (not PDF but XML, CSV, etc.), and can be re-used without restriction in the form of cost, compulsory registration, etc. (Algemene Rekenkamer, 2014).

The Ministry of Education, Culture and Science (OCW) generates much education data that fits the definition of open public data which are interesting for the Dutch citizens. For example, journalists who want to know more about the school performance in the Netherlands, parents who want to know the characteristics of schools before enrolling their children, and businesses that want to know the prospect of hiring skilled employees. OCW describes the move as education quality openness, with the future vision as follow (Rijksoverheid, 2018b):

- Parents and pupils know where they can find important information about schools.
- Parents and pupils use this information to compare schools and choose a suitable school.
- Parents, pupils, and the education council use the information to discuss the quality of education with the school.
- All schools use the available data in the best possible way to improve education.
- All government data is public and is used to develop useful applications for parents, students, teachers, and school leaders.

Previous experience in the openness and transparency

The Education Executive Agency (DUO) is an implementing agency which collects, manage, and enriches educational data on behalf of OCW. In the past, DUO received and processed many public data access (WOB) requests from the public about the education data. For example, the request for school performance data by newspaper Trouw in 1997. It led to increased transparency from government authority where they proactively published the education datasets (openstate.eu, 2016).

This demand for openness initiates the project “Windows for Accountability”, where the schools actively publish the data about themselves that will help parents to choose a suitable school for the children, by VO and PO Raad (primary and secondary education school association) (Hanne Obbink, 2012). One of the results is the scholenopdekaart.nl website which uses data from DUO, Inspectorate of Education, and the data provided by the schools. Using this website parents, pupils, and other interested parties could compare schools and gain insight about the education quality in their area (scholenopdekaart.nl, 2018). This initiative started even before the nation-wide commitment for open government action plan in 2013. The scholenopdekaart.nl project for secondary school last from 2007-2013 and the subsequent project for primary school from 2012-2016. However, only 16% of parents who search for schools information on the internet use the website in 2015 (Ministry of Education, 2015).

Current open education data policy

OCW publish its open data in several portals such as duo.nl, onderwijsinspectie.nl, and onderwijsincijfers.nl. All these data are also registered in the national open data portal called data.overheid.nl. Each of the portals has a different type of data generated by the respective government agency (DUO, Education Inspection Agency, and OCW).

Table 2 List of OCW Open Data Portals and the contents

Organization	Data Portal	Information
Education Executive Agency (DUO)	www.duo.nl/open_onderwijsdata/	• Education datasets for the diverse level of education such as Primair Onderwijs (PO), Voortgezet Onderwijs (VO), Middelbaar Beroepsonderwijs (MBO), and Hoger Onderwijs (HO). It contains school-related information (school address, school status), students related information (number of students), staff related information (number of the teacher), and school funding data. • Prognosis data (the education year 2017-2036) for the number of students in primary and secondary education to help schools in planning their budget and facilities. • Education Data API (last update 2016) which has 93 datasets from primary and secondary education
Education Inspection Agency	www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata	• Data about the indicators and standards used in the school assessments. • Results of the school assessments. • Sample data of schools that are selected for assessments.
Ministry of Education, Culture and Science (OCW)	www.onderwijsincijfers.nl www.ocwincijfers.nl www.trendsinbeeldocw.nl	• Onderwijs in cijfers (Education Data in figures) is the collaboration between DUO, OCW, and Central Bureau of Statistics (CBS) which present national education data in figures. • OCW in cijfers (OCW in figures) is the website that presents the figures from all section of OCW (education, culture and media, science and emancipation). • 'Trends in beeld' is the website that provides the monitoring information of OCW such as policy agenda, policy target, and budgeting for each section of OCW.

Furthermore, OCW arranges events for open education data which brings parents, students, teachers and school management together to discuss the possible application of open data. It is arranged in November 2016 with the theme "Education Data under scrutiny". In this event, the participants came with several ideas to utilize open data which lead to one question as a use case "How can I make a good secondary school choice based on my values?" (Rijksoverheid, 2016). Afterward, the event hosted a hackathon to create application prototype that answers the use case. In 2018, OCW and municipality of Amsterdam organize a hackathon "Hack de Valse Start" to address the question of inequality opportunities in the education (openstate.eu, 2018). This hackathon aims to gain insight into unequal opportunities by combining education and municipality open data provided by DUO and Central Bureau of Statistics.

On 25th May of 2018, the General Data Protection Regulation (GDPR) is formally applied in the Netherlands. The introduction of GDPR reinforces the existing barrier faced by government agencies in opening their data (risk-averse culture and limited resource to handle the data publishing process). The risk of opening data is increased because there is a hefty fine in case of data breaches (as high as €20 million or €10 million according to the bill).

In order to comply with the data protection specification of the GDPR, sizeable resources are required (both human resources and monetary) which will put pressure on their budget for other functions. OCW hires two Data Protection Officers, one at DUO and one at the board department. A specific FG at DUO was chosen because of the vast amount of personal data at DUO and the need to exercise adequate supervision at a short distance (OCW, 2017). The Data Protection Officer is in charge of Data Protection Impact Assessments (DPIA), mapping the privacy risks of a data processing system in advance and take measures to reduce the risks.

For example, DUO requires the amount of €12 million in 2018, increasing to €27 million in 2022 to implement the changes required by GDPR; government concludes that with the existing problems in OCW budget no room for this expenditure within the 2018 budget (OCW, 2017).

The existing open data policy focus on the data stewardship capability to ensure the supply of open data. However, there are limited functional applications resulted from the open datasets such as scholenopdekaart.nl and studiekeuze123.nl. The introduction of GDPR also creates another pressure for the government agencies in charge of open education data (OCW, DUO, and Education Inspection Agency). Given that each of this agency has limited (personnel and monetary) resources, how do they select the attributes for the open education data policy and justify the choices being made?

1.2. Problem Definition

The data provisioning model, drivers, and barriers

Sieber & Johnson (2015) describe four open data provision models based on the interaction between government and citizen as shown in Figure 2.

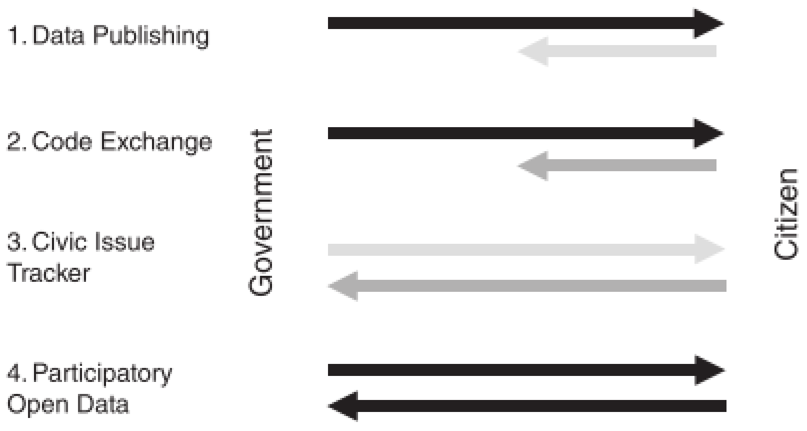


Figure 2 Models of open data provision adapted from (Sieber & Johnson, 2015)*. Darker lines represent more significant interactions.*

The Netherlands provides the data in a unidirectional way, which means that the data is provided one-directional from the data owner (e.g., government, or potential non-profit organization) to the end user or developer (citizen, community organization, or private sector) with minimum feedback from these end users or developers.

Janssen, Charalabidis, & Zuiderwijk (2012) found that contributing to public value creation (transparency and accountability) and economic growth are the main drivers for open data initiatives in the Netherlands. For example, one of the respondents believed that by opening government data citizen could confirm and verify whether policymaker's conclusions are correct and justified. Opening data in itself are seen as an altruistic act which can enhance the public's perception of government transparency; if the government does not have anything to hide it will share its data openly and freely. Furthermore, the potential economic growth benefits are based on the prospect of using open government data to create innovative services and inform potential investors and companies.

However, those potential benefits can only be attained after addressing several barriers which are commonly categorized as data provider and data use barriers (Janssen et al., 2012). The data provider barriers encourage a restrained attitude of the data provider in publishing data while the data user barriers affect the usability of open data.

Examples of data provider barriers are institutional barriers (risk-averse culture, limited resources to handle the data publishing process), whereas the data use barriers are task complexity (lack of metadata, lack of skills to discover the data) and use and participation (lack of incentives, insufficient knowledge to process the data).

One of the barriers is 'unclear trade-off between public values for the policymaker'. For example, should policymakers rigorously pursue transparency with the risk of compromising the privacy value? How can the policymaker justify the pursuit of a specific value over the others?

The Dutch open data policy characteristics

The Dutch open data policies objective is to create public values such as transparency, economic growth, and innovation (Ministry of the Interior and Kingdom Relations, 2015). However, existing performance indicators are focused on the publishing process of the data and how to deal with the risk associated with it (confidentiality, privacy, data quality, completeness, misuse and misinterpretation) (Ministry of the Interior and Kingdom Relations, 2017).

Zuiderwijk & Janssen (2014) compared seven Dutch governmental policies and found several characteristics of the Dutch open data policy such as lack of systematic collaboration and 'jumping on the bandwagon' tendency. The government agencies are susceptible to mimic the other agencies that it deemed successful and followed their "best practice" regardless of their data context and the environment they are operating in. This tendency is not exclusive to the Dutch open data policy as shown in the study by Zuiderwijk, Shinde, & Janssen (2018).

Zuiderwijk, Shinde, & Janssen (2018) identified a mismatch between open data policy objectives and the actual benefits derived from those policies based on 168 survey responses concerning 156 open government data initiatives at different government levels worldwide. Their study shows that there is no statistically significant relation between the policy objectives of Open Government Data Initiatives (OGDI) and its delivered benefits. For example, there is no significant difference in the delivered benefit "easier access to data" between OGDI that stated openness as its policy objectives compared to a policy that does not state it.

The finding shows that practitioners tendency to mimic other initiatives might lead them to overlook the objectives, the context and the deliverance of societal values which are unique to the domain they operate.

Consequently, governments choose to measure the open data policy performance based on the straightforward attributes such as the quantity of the data published and scores from the international benchmarks (Algemene Rekenkamer, 2016; data.overheid.nl, n.d.). Using the quantity of the data published have its limitation because either the data is useful or not for the users, it is still counted in the sum. It gives the data providers the impression that they have achieved something even though the published data are not being used to create the desired public values. The benchmarks were developed to serve different purposes with varying degree of specificity, scope, and focus (Susha, Zuiderwijk, Janssen, & Grönlund, 2015). Applying it without discrimination produce results that are generic and ambiguous for any particular organization.

Susha et al. (2015) concluded that the creation of model and benchmarks should be guided from the perspective of what is beneficial for open data end users since it is the primary goal of opening data. The policymakers tendency to mimic each other and settle for generic performance indicators (quantity and benchmark scores) show their lack of insight into the citizen preferences for open data policy attributes.

Problem 1: Lack of insight into the citizen preferences of open data policy attributes

In order to address the problem, this research aims to identify the citizens preferences for the open data policy through the citizens stated choice experiment (CSCE). The citizen stated choice experiment is a type of discrete choice experiment (DCE). The discrete choice experiment is a quantitative technique to elicit individual preferences (Mangham, Hanson, & McPake, 2009). The researcher presents several ‘hypothetical’ alternatives and infers how individuals value selected attributes of programs, products, services, or policies based on their choices.

Using citizen stated choice experiment to infer citizen preferences

Two dominant valuation methods are revealed preference and stated preference method (DCE is categorized as a stated preference method). Revealed preference methods assume that actual preferences from individuals can be derived from direct observations and responses from individuals to complement or substitute goods (Cook, Davídsdóttir, & Kristófersson, 2016). The gathered data is based on what an individual did in a specific situation. However, in many public goods such as environmental valuation, human health effects, and other outcomes for which (direct or indirect) revealed preference (RP) data are not available; stated preference methods are the only known approach to estimate values for changes (Johnston et al., 2017). In comparison to other stated preference techniques that require the individual to rank or rate alternatives (ranking and best-worst scaling), a DCE presents a reasonably straightforward task and one which more closely resembles a real-world decision (Mangham et al., 2009).

The main critique for stated preference method is whether SP methods can provide credible information to inform decision-making because the respondents are asked to choose between ‘hypothetical alternatives’. Particular attention has been given to the issue of hypothetical bias, or whether values estimated using SP data are equivalent to those that would be estimated using parallel RP data (in cases where valid comparisons are possible) (Johnston et al., 2017).

Currently, no study investigate the valuation of open data policy attributes, either the benefits (e.g., transparency, participation, openness, engagement, economic gain) or risks (privacy breach, misuse and misinformation) from a citizen perspective and how it can be used for the policy decision making process.

Problem 2: Lack of study that empirically assesses open data policy attributes from the citizen perspectives

The citizen preferences and measured trade-off attributes are important components to understand the existing gap between open data policy objectives and the realized benefits. In the current process, policymaker measures the policy performance from the data provider perspective. This research could lead to a new performance indicator based on the citizens needs that the policy accommodates and how the citizens perceived the fulfillment. Other than that, identifying the citizen preferences in the agenda settings phase will help policymaker to accurately allocate their resource according to the citizen's needs and prevent futile implementation. Furthermore, policymaker can create a citizen-informed decision making when they deal with various policy alternatives.

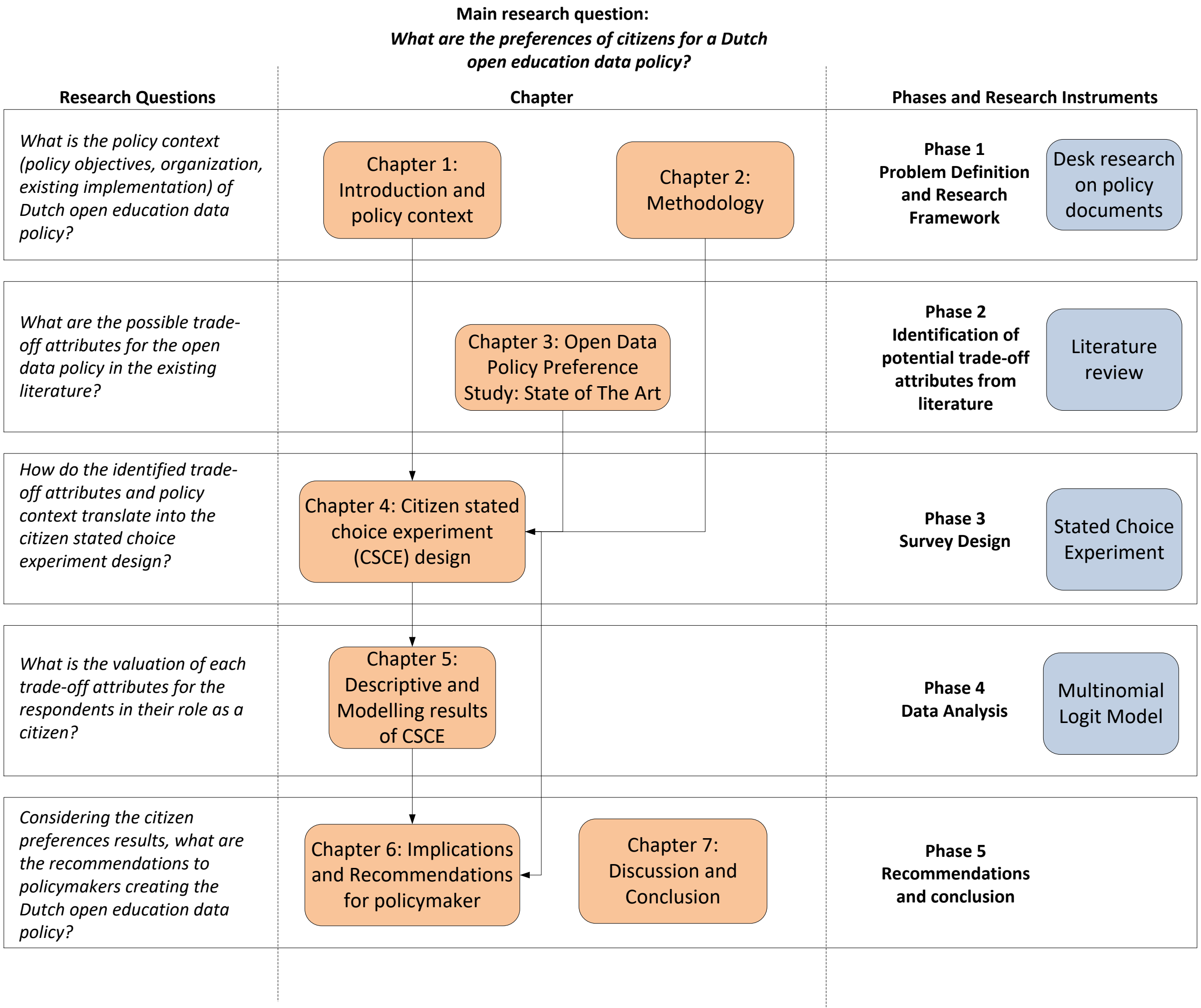
Therefore, the ultimate aim of this research is to empirically measure citizen preferences of open data policy attributes specifically the Dutch open education data. The results are meant to identify the relative preferences of the open education data policy attributes from the citizen perspective and how policymaker can utilize it to create a suitable open education data policy.

The Dutch open education data is chosen because the domain has long experience in openness and transparency of public data. Starting from publishing school performance data in 1997 and implementing the project "Windows for Accountability" in 2012, before the nation-wide commitment for open government action plan in 2013. Furthermore, higher education students as the target respondents are more likely to have experience interacting with open data in the education domain which will improve their comprehensibility of the stated choice experiment.

1.3. Research Design

Based on the contextual background and scientific gap explained in section 1.2. Problem Definition, the main research question for this study is:

What are the preferences of citizens for a Dutch open education data policy?



Main research question:
What are the preferences of citizens for a Dutch open education data policy?

Research Questions	Chapter	Phases and Research Instruments
What is the policy context (policy objectives, organization, existing implementation) of Dutch open education data policy?	Chapter 1: Introduction and policy context Chapter 2: Methodology	**Phase 1 Problem Definition and Research Framework** Desk research on policy documents
What are the possible trade-off attributes for the open data policy in the existing literature?	Chapter 3: Open Data Policy Preference Study: State of The Art	**Phase 2 Identification of potential trade-off attributes from literature** Literature review
How do the identified trade-off attributes and policy context translate into the citizen stated choice experiment design?	Chapter 4: Citizen stated choice experiment (CSCE) design	**Phase 3 Survey Design** Stated Choice Experiment
What is the valuation of each trade-off attributes for the respondents in their role as a citizen?	Chapter 5: Descriptive and Modelling results of CSCE	**Phase 4 Data Analysis** Multinomial Logit Model
Considering the citizen preferences results, what are the recommendations to policymakers creating the Dutch open education data policy?	Chapter 6: Implications and Recommendations for policymaker Chapter 7: Discussion and Conclusion	**Phase 5 Recommendations and conclusion**

Figure 3 Research design

In order to answer the main research question, the following research questions are derived:

1. What is the policy context (policy objectives, organization, existing implementation) of Dutch open education data policy?

It is essential to understand the context of Dutch open education data policy to answer the main research question. In the desk research, the **policy objectives**, **organization context**, and **existing implementation** of Dutch open education data policy is explored.

The desk research is conducted using policy documents published by Ministry of Education, Culture and Science (OCW), Education Executive Agency (DUO), and Education Inspection Agency (Inspectie van het Onderwijs) which are the government agencies in charge of providing open education data.

2. What are the possible trade-off attributes for the open data policy in the existing literature?

   In order to answer the main research question, a stated choice experiment is conducted using the survey as the media. The survey design requires a set of trade-off attributes to construct policy alternatives; the respondents' preferences are then inferred from their choice of policy.

   A literature review is conducted to identify possible trade-off attributes from the existing open data policy literature. The result of the literature review is a list of potential trade-off attributes for survey design.

3. How do the identified trade-off attributes and policy context translate into the citizen stated choice experiment design?

   The citizen stated choice experiment is designed in the form of a survey with narratives and choice tasks. The information from RQ1 (the specific policy context of Dutch open education data) and RQ2 (potential trade-off attributes from literature) is combined for the survey design.

   The narratives incorporated the information about organization context, policy objectives, and existing implementation to ensure the realism of the experiment. Furthermore, the trade-off attributes are used to formulate alternatives for the choice tasks. A pilot study is conducted with limited respondents to test the survey in order to create a realistic and relevant final survey.

4. What is the valuation of each trade-off attributes for the respondents in their role as a citizen?

   The final survey is distributed among Dutch higher education students. In the citizen stated choice experiment, citizens preferences can be inferred from their valuation of each trade-off attributes. Multinomial Logit (MNL) model is generated to analyze the collected observations. The result of the MNL model is an estimation of each trade-off attribute value/utility for the citizens.

5. Considering the citizen preferences results, what are the recommendations to policymakers creating the Dutch open education data policy?

   The result of citizens stated choice experiment and its implication for the open education data policy are discussed. For example, policymakers can design a policy that maximizes the value/utility for the citizen or comparing different policy alternatives from the citizen perspectives.

Research Flow

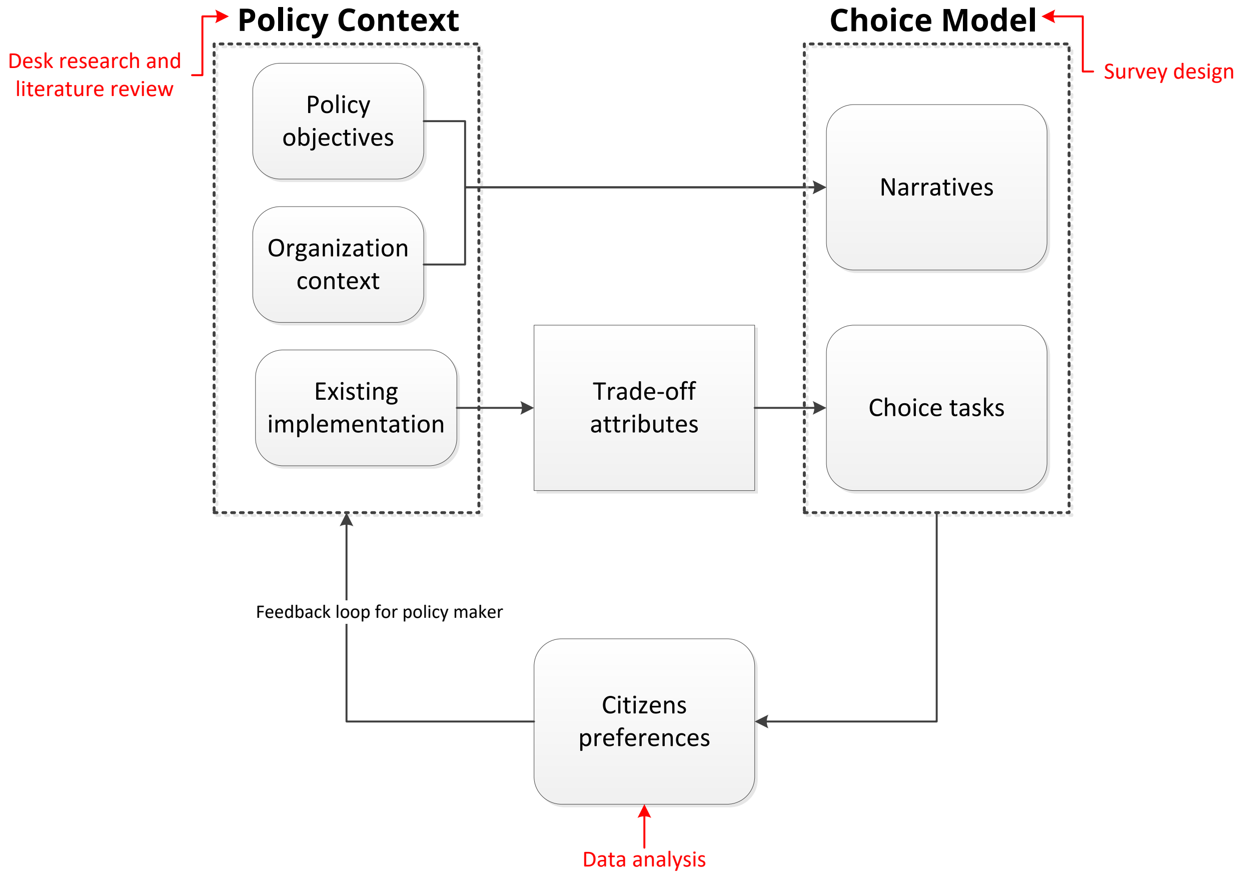


Figure 4 Research framework and method, adapted from (Johnston et al., 2017; Mangham et al., 2009)

The overall research framework can be seen in Figure 4. The following paragraphs will describe the research activities and the methods/tools used.

First, the open data policy context is established through **literature review** and **desk research**. In the literature review, existing open data policy studies are explored to identify possible trade-off attributes between different open data policies.

The resulting attributes from the literature review are then combined with information from desk research on existing policy documents of related government agencies. Furthermore, the desk research over the policy document is important to understand the organizational context of the policy. For example, open data policy in the organizations which handle private data is more concerned with the confidentiality and the privacy protection of the published data compared to the organization which handles geographical or public property data.

Afterward, the results of the policy context analysis are incorporated in the **survey design** of citizens stated choice experiment. The organization context and policy objectives are used for the introductory passage and leading questions before the choice tasks to provide the realism and precisely delineate the tasks for the respondents. Several studies (Carson & Groves, 2007; Johnston et al., 2017) emphasize the importance of consequentiality in designing the stated choice experiment. Consequentiality means the respondents perceive that their answers are potentially influencing the government's actions. The survey design will follow guidelines based on best practice stated preference studies (Arrow et al., 1993; Johnston et al., 2017). For example, referendum format where the implied decision mechanism for a policy to be implemented is majority vote (Arrow et al., 1993) and arranging the choice task in a single binary question which represents the baseline (status quo) and the proposed alternatives (Johnston et al., 2017).

A **statistical model** is created to analyze the choices made by the respondents in the survey. Random utility theory by McFadden (1974) is the established approach to relate the deterministic model with a statistical model of human behavior. The previous citizen stated choice experiment studies (Mouter & Chorus, 2016; Mouter, van Cranenburgh, & van Wee, 2017a, 2017b) employed Multinomial Logit Model (MNL), Mixed Logit Model, and Latent Class Analysis (LCA) to measure the citizen preferences.

MNL focus on the average preferences which result in parsimonious estimator with a unique solution; it also requires the smallest sample size compared to Mixed Logit Model and LCA (Hauber et al., 2016). However, this simplistic approach means that the MNL model assumes homogeneity in preferences among the respondents and does not account for panel nature of the data.

Mixed Logit Model able to capture the model heterogeneity and accounts for the panel nature of the data; Mixed Logit Model explicitly assumes that there is a distribution of preference weights across the sample reflecting differences in preferences among respondents, and it models the parameters of that distribution for each attribute level (Hauber et al., 2016). However, there is little direct guidance available to determine the appropriate functional form for the distribution of preferences across respondents. The mixed logit model is more difficult to use than MNL and requires larger sample sizes than MNL. It also requires assumptions about the distribution of parameters across respondents which are difficult to determine a priori because individual preference weights are not directly interpretable.

Latent Class Analysis (LCA) also able to model the heterogeneity using latent classes which result in parsimonious estimator with a unique solution; it requires smaller samples than Mixed Logit Model (Hauber et al., 2016). However, it requires the assumption to determine an appropriate number of classes to be estimated, and the required sample size varies with the number of classes in the model. LCA is also difficult to interpret when the chance of being in all classes is more or less the same across respondents.

This research will use MNL approach due to its simplicity and comprehensibility for a first attempt to empirically measure citizen preference for open data policy. Train (2003) describes MNL ability to represent systematic 'taste' variations (i.e., those related to observed characteristics of the respondents) which are valuable for descriptive results of the model. The statistical model analysis will result in the statistical measurement of citizen preferences over the trade-off attributes and the descriptive results of the model. The citizen preferences result is communicated to the policymaker in the form of implications and recommendations for their decision-making. For example, to what extent their existing open data policy reflect the citizen preferences, how they can develop an open data policy which better reflects citizen preferences. Citizens are the end user of open data and knowing their preferences helps the policymaker to decide on the open data policy that maximizes the utility derived by the citizens.

Data Requirements

The following data are required to conduct the research:

Table 3 Data requirements

Data	Sources
Policy objectives	• Literature review: ○ Existing studies on the open government data preference study • Desk research: ○ Open data policy documents from the Dutch education agency. All the formal documents can be found in rijksoverheid.nl (government portal which publishes policy plan, laws, and regulation)
Organization context	
Policy alternatives	
Citizens preferences	• Self-distributed survey (TU Delft students and their network of friends)

The self-distributed survey is conducted using an online survey tool (SurveyGizmo). The online survey tool provides flexibility for the distribution of the survey (URL, QR code) and compiling the collected observations. It allows different question formats (multiple checkboxes, radio buttons, Likert scale, textbox) and the inclusion of media (image, audio, video).

The survey results will be processed using data analysis tools (Python Biogeme) an open source freeware designed for the maximum likelihood estimation of parametric models in general, with a special emphasis on discrete choice models (Bierlaire, 2016). It provides the required package to analyze Multinomial Logit Model (MNL), and the online documentation are widely available. It is also the software used by the previous citizen stated choice experiment (Mouter & Chorus, 2016; Mouter et al., 2017a, 2017b).

Scientific and social contribution

The citizen stated choice experiment had been used in the domain of transport policy to assess citizen preferences for spatial equality in the context of Dutch national transport plan (Mouter et al., 2017a) and individuals trade-off safety and travel time in their role as a citizen (Mouter et al., 2017b). The discrete choice experiment is also commonly used for the valuation of non-market goods in the environmental policy (Achtnicht, 2011; Cook et al., 2016) and patient preference analysis in the health policy settings (Cheraghi-Sohi et al., 2008; Rubin, Bate, & George, 2006; Ryan, 2004).

This research is a first attempt to extend Mouter & Chorus' (2016) citizen stated choice experiment approach for the valuation of citizen preferences in the context of open education data policy. The experiment will infer the individuals preferences in their role as a citizen and measure the trade-off that they make in considering different open education data policy alternatives.

Previous study that assess the Dutch open data policy mostly use qualitative approach such as literature review and interviews (Welle Donker & van Loenen, 2017), deep interview (Westra & Poel, 2017), and case study with a social cost-benefit analysis (Welle Donker, van Loenen, & Korthals Altes, 2017). The research proposed in this study contributes to the utilization of a quantitative approach to assess citizen preferences in the Dutch open data policy (specifically citizen stated choice experiment).

There is a previous attempt to empirically measures the performance of government open data websites and the acceptance and use of these data from a citizen perspective in the United Kingdom (UK) by Weerakkody, Irani, Kapoor, Sivarajah, & Dwivedi (2017) using an adjusted diffusion of innovation model as predictors. The research proposed in this study contributes to the emerging needs to empirically assess the open data policy from citizens perspectives (their preferences and perceptions).

This research provides an alternative method for governments to evaluate and develop their open data policy alongside the commonly used government/data provider perspective. It enables policymaker to empirically valuate citizen preferences for 'hypothetical' open data policy and develops suitable open data policy from the citizen perspective. Obtaining this insight will help policymakers to see the open data policy from the eye of open data end users (the primary beneficiary of opening data) and bridge the existing gap between open data policy objectives and the realized benefits.

The valuation enables policymakers to understand how citizens valuate specific attributes in comparison to the others and what is the trade-off for the policymaker if they choose one attribute over the other. For example, if the policymaker knows the relative value of organizing a participation and engagement events (e.g., hackathon) compared to commissioning the creation of service from open data (e.g., studiekeuze123.nl); they can weight the trade-off in their decision-making process.

1.4. Thesis Structure

The thesis is arranged into five phases as shown in Figure 3:

1. Problem definition and research framework
2. Identification of potential trade-off attributes from literature
3. Survey design
4. Data analysis
5. Recommendations & conclusion

Chapter 1 comprises of the problem definition and the research questions, in this part the context of Dutch open education data policy is investigated from the policy documents, and the research questions are formulated based on the identified problem. Chapter 2 discuss the background theory and the chosen methods to conduct the research. Chapter 3 explores the state of the art of the open data policy study through a literature review. In Chapter 4, the detail of survey design and choices made (e.g., selection of attributes, narratives building) are discussed. Chapter 5 explicate and discuss the descriptive result (e.g., the sample characteristics and their choice behavior) and modeling result (the MNL model estimation of citizen preferences) of the survey. Chapter 6 will discuss the implication of descriptive results and the model estimation of the trade-off attributes for the open data policy including the recommendations for the policymaker. Finally, Chapter 7 reflects on the whole research process and provides the research conclusion.

Chapter 2: Methodology

In this chapter, the methodology for the research is explained. The rationale for choosing citizen stated choice experiment over the consumer stated choice experiment is discussed. After that, this chapter describes the step-by-step approach to design a Citizen Stated Choice Experiment and the best practices found in the discrete choice experiment literature. Finally, the choices being made are summarized in conclusion.

1.1. Why a Citizen Stated Choice Experiment (CSCE)?

The discrete choice experiment is a quantitative technique to elicit individual preferences (Mangham et al., 2009). In DCE, researchers present several hypothetical alternatives and infer how individuals value selected attributes of programs, products, services, or policies based on their choices.

DCE has been applied in health policy settings for different cases such as resource allocations, patient's priority in the care services, and their policy choices regarding access to a general practitioner (Cheraghi-Sohi et al., 2008; Rubin et al., 2006; Ryan, 2004). Environmental policy to value non-market environmental goods in the decision-making process, for example, the impact of environmental benefits (lower CO2 emission) in the house owner decision making for the choices of heating (Achtnicht, 2011). In the transport policy, DCE has been applied to assess individuals' preference and how they value trade-off attributes (travel time and safety, spatial equality) between different transport policies (Mouter et al., 2017a, 2017b).

The citizen stated choice experiment is a variant of the discrete choice experiment (DCE) where the respondents are asked to do the choice tasks from 'citizen' perspective instead of 'consumer' perspective.

Mouter & Chorus (2016) differentiate between consumer and citizen perspective based on different budget constraints. Consumer preference if the choice involves the after-tax income of the individual; and citizen preference if the choice is based on previously collected tax by the government. In the study, Mouter & Chorus (2016) empirically confirm the difference between the individual valuation of time gained depending on their role as 'consumer' or 'citizen'. Further research by Mouter et al. (2017a, 2017b) extend the notion of citizen stated choice experiment for other non-market goods valuation in transport policy such as safety and spatial equality.

The identification of preferences from citizen perspectives is suitable for open data policy because the provision of open data is fully-funded from the government budget. The Dutch government also stated that open public data can be re-used without restriction in the form of cost, compulsory registration, etc. (Algemene Rekenkamer, 2014). Therefore, there is no scenario for consumer preferences in the current open data policy context, including consumer preferences will affect the realism of the study because the respondents do not have any experience/baseline information about paid open data.

1.2. Approach to create Citizen Stated Choice Experiment (CSCE)

There are several phases of designing a choice experiment which is summarized in Table 4. The following section will explain the activities in each phase and its best practices.

Table 4 Designing Discrete Choice Experiment

Phases (based on (Mangham et al., 2009))	**Adaptation to this research**	**Relevant section**
Establishing attributes	• A literature review of existing open data policy assessment study • Desk research on existing policy documents	Chapter 1 and Chapter 3
Assigning attribute levels	• Desk research on existing policy documents	Chapter 1 and Chapter 4
Designing the choice sets	• Create balanced and orthogonal survey design	Chapter 4
Generating, pre-testing, and distributing the questionnaire	• Pilot test questionnaire • Revised the questionnaire based on the input from the pilot test • Distribute the final questionnaire	Chapter 4
Analyzing DCE data	• Create a statistical model to analyze questionnaire results. • Explain the result of the statistical model and its implication for the policymaker	Chapter 5 and Chapter 6

CSCE phase 1: Establishing attributes

In this phase, the researcher identifies relevant attributes for the stated research question (e.g., in this study the open data policy). This activity requires a good understanding of the target population and perspective. Other than that, policy concerns from the local institutions and policymaker can be used to identify the attributes. All of this information can be obtained from published and grey literature such as previous study, policy documents, and government reports.

In choosing the attributes, there is a need to balance statistical efficiency and respondents' cognitive capacity (or response efficiency) (Johnston et al., 2017). In practice, most DCEs selected less than ten attributes to ensure respondents ability to compare all attributes listed when making their choice (DeShazo & Fermo, 2002). Designs should include a limited number of attributes that are particularly relevant to decision-makers and respondents. Great number of attributes could encourage participants to develop simple decision rules where they choose based on a single or subset of attributes (Mangham et al., 2009). Other than that, it is important to avoid inter-attribute correlation, the conceptual overlap between two or more variable, since it would impact the accurate estimation of a single attribute effect toward the dependent variable (Mangham et al., 2009).

CSCE phase 2: Assigning attribute levels

After all of the attributes are established, the researcher needs to assign the attribute levels that reflect the range of situations that the respondent might expect to experience (Mangham et al., 2009). This information can be obtained from the pre-testing with potential respondent, during this interaction researcher can ask whether the assigned levels are realistic or not. Ensuring the levels are realistic and meaningful will increase the precision of parameter estimates (Hall, Viney, Haas, & Louviere, 2004). Other than that, the levels assignment should follow the utility functions being used, for example when the change in one attribute is linear and the other is non-linear, this difference should be reflected in the levels of attributes (Johnston et al., 2017). In this step, the base level of attributes is established from the status quo (current condition) and an additional level which reflects the reasonable improvement from status quo (Mangham et al., 2009).

CSCE phase 3: Designing the choice sets

The next phase is combining the attributes and the assigned level to create choice sets which reflect the hypothetical alternatives. A full factorial design that combines all the possible attributes and attribute levels will be able to estimate the main effects and interaction effect of all the attributes. The direct effect is the changes in respondent choice based on the variance in an attribute levels (e.g., difference in budget) while the interaction effect is the changes in respondent choice due to the combination of two or more attribute levels together (e.g., difference in budget combined with difference in size) (Mangham et al., 2009).

However, it would be too cost-prohibitive and tedious task for respondents to finish the choice tasks of a full factorial design (Kuhfeld, 2010). Therefore, the researcher chooses a fractional factorial design in the selection of possible alternatives. The fractional factorial design should aim for a balanced and orthogonal design. The orthogonal design is achieved when the parameter estimates in the linear model are uncorrelated; each attribute is statistically independent of the others. The balanced design is achieved when each attribute level occurs equally often, which minimizes the variance in the parameter estimates. However, there will be a trade-off between orthogonal and balanced design, and the researcher can select the most efficient design using a measure known as D-efficiency (Kuhfeld, 2010).

CSCE phase 4: Generating, pre-testing, and distribute the questionnaire

The designed choice sets become the basis for the alternatives presented in the questionnaire. The number of choice sets presented to respondent depends on the size of the fractional design. Other than that, the researcher should consider the boredom threshold, a practical limit of how many choice tasks can be completed by the respondent before the boredom sets in. This boredom threshold will depend on the number of choice sets, its complexity, and target population (Mangham et al., 2009).

A pairwise design where the respondents are asked to consider a choice set with two alternatives and stated their preference could represent the demand conditional on accepting one of two scenarios (Mangham et al., 2009). However, a researcher could introduce non-choice demand by presenting “choose neither” option which allowed the respondent to reject both alternatives and provide data to estimate actual demand.

The choice tasks should be presented in a randomized order to avoid information order effect (Kjær & Gyrd-Hansen, 2008). The questionnaire should be clearly presented and contain a standard introduction to the DCE with choice set examples, pictures, diagram, and symbols may improve the comprehensibility of the questionnaire (Mangham et al., 2009). The questionnaire should also collect socio-economic indicator of the respondent to analyze the impact of individual characteristics on the choices made.

The questionnaire is pre-tested to limited respondent and researcher could review several elements of the design process. In the pre-testing researcher can validate the selection and definition of attributes and their levels (Hall et al., 2004). The researcher can ask the respondent whether there is conceptual overlap between attributes, does the attribute levels represent the reality, or does the wording of the attributes create a biased view on one of the choices. Furthermore, the pre-testing should check the respondent's understanding of the task, their ease of comprehension and whether the number of choice sets can be managed by the target population (Hall et al., 2004).

CSCE phase 5: Analyze DCE data

The analysis of DCE data are based on the combination of two theory: 1) Lancaster's characteristic demand theory (Lancaster, 1966), and 2) Random Utility Theory (McFadden, 1974). The characteristic demand theory describes that consumer derived utility from the characteristics of goods rather than the consumption of the goods itself. This approach allows us to infer individual preferences based on their choice of characteristics (attributes) presented in the options.

The random utility theory allows the researcher to analyze the utility derived from the characteristics of the goods. The amount of utility is represented by a relative and abstract numerical value, while choices are the only visible indicator of utility. Individuals expressed their preference from the amount of utility that they perceived, satisfaction when the specific attributes provide a positive utility and dissatisfaction for a negative utility. Other than that, the analysis is conducted on the basis that every individual is rationally maximizing utility who chooses an alternative that gives the largest relative utility. The utility function can use the linear or non-linear parameter. In its simplest form, the utility function can be defined as a linear expression in which each attribute is weighted by a unique parameter to account for that attribute's marginal utility (Mangham et al., 2009).

Key-elements of RUM-choice model:

- i, j = alternatives in the choice sets (i is alternative 1 and j is alternative 2)
- m = attributes (e.g., cost, time)
- X = attribute values from observation
- β = parameters to be estimated
- ε = randomness (all the unobserved determinants of the utility)

The systematic utility (V_i) is the utility that can be related to observed factors (e.g., cost, time, age, income level) which can be represented in the form of:

$$V_i = \sum_m \beta_m \cdot X_{im}$$

The total utility of an alternative can be represented through this equation:

$$U_i = V_i + \varepsilon_i = \sum_m \beta_m \cdot X_{im} + \varepsilon_i$$

An alternative is chosen if its total utility is the largest. Therefore, alternative i will be chosen over alternative j if it fulfills this condition:

$$\sum_m \beta_m \cdot X_{im} + \varepsilon_i > \sum_m \beta_m \cdot X_{jm} + \varepsilon_j, \forall j \neq i$$

The total utility (U_i) is the combination of systematic utility (V_i) and error term (ε_i). There will be a situation where an individual does not choose an alternative with the highest systematic utility due to the unobserved factors from error term. It implies that the prediction of choices is based on probability with an assumption (higher systematic utility → higher choice probability). The probability of alternative i is chosen over alternative j can be expressed as follow:

$$P(i) = P(V_i + \varepsilon_i > V_j + \varepsilon_j, \forall j \neq i)$$

It is important to note that the utility is not an absolute value. Therefore, what matters in the choice situation between alternative i and j is the utility differences of alternative i relative to alternative j. The probability equation can be rewritten as:

$$P(i) = P(U_i - U_j > 0, \forall j \neq i)$$

In this research, the multinomial logit model (MNL) is used to explicitly estimate the β which is the parameter that determines the individual difference/taste for a certain attribute/characteristic. The probability equation of choosing alternative i, if ε ~ EV Type 1 with variance $\pi^2/6$ in MNL model can be written as follow:

$$P(i) = P(V_i + \varepsilon_i > V_j + \varepsilon_j, \forall j \neq i) = \frac{e^{V_i}}{\sum_{j=1\ldots J} e^{V_j}} = \frac{e^{\sum_m \beta_m X_{im}}}{\sum_{j=1\ldots J} e^{\sum_m \beta_m X_{jm}}}$$

(Note: in the denominator, *J* denotes choice set size. *j* runs from 1 to , and includes *i*)

Estimating β implies inferring the importance of the attribute (e.g., cost) relative to another observed attribute (e.g., time) and relative to unobserved factors (‘randomness/error term’).

1.3. Conclusion

The study exclusively select citizen stated choice experiment over the consumer stated choice experiment due to the reality of open data policy implementation which is fully-funded from the government budget and provided without any charge for its utilization. Five steps of conducting the discrete choice experiment are explained with the best practices of each step. The operationalization of these steps will be covered in Chapter 3 until Chapter 5 of this report.

This study based the selection of the attributes on three criteria: the expected influence on an individual, the societal relevance of the factor, and measurability in the discrete choice experiment. The orthogonal design is favored over D-efficient design due to its robustness for an experiment without established prior values (estimated parameters found from the previous study). The D-efficient design is the most efficient for an experiment if the prior values used to generate the design is accurate and close to the true values, however it is risky to use D-efficient design with uncertain prior because the design significantly becomes inefficient when the true values deviate from the prior values (Walker, Wang, Thorhauge, & Ben-Akiva, 2018).

The analysis of the citizen stated choice experiment is conducted based on the Lancaster's characteristic demand theory (Lancaster, 1966) and Random Utility Theory (McFadden, 1974). Combining both theories enable the research to explicitly estimate citizen preference of attributes (β) based on the alternative chosen by a citizen in the designed choice situations.

Chapter 3: Open Data Policy Preference Study: State of The Art

This chapter is aimed to answer sub research question 2: *What are the possible trade-off attributes for the open data policy in the existing literature?*

In this chapter, we investigate the state of the art of open data policy preference studies. First, the literature review search strategy is discussed. Next, the types of open government data study and the reason behind the diversity are discussed. Afterward, this chapter discusses the extent of existing preference studies and the perspectives that it takes in the study. Finally, the summary of recurring attributes described in the existing open government data study is presented.

3.1. Literature review search strategy

The literature review is conducted through SCOPUS using the terms 'open government data', 'preference'. The terms 'measurement', 'assessment' and 'evaluation' are chosen to extend the scope of literature because using 'preference' term result in a limited number of literature (24 journal papers). In the measurement, assessment, and evaluation study, the open data policy is scrutinized from different perspectives and aspects which are suitable to identify open data policy attributes.

The source type is limited to journal with the topic of social science and publication year between 2013-2018. The topic is limited to social science because there is a similar study in computer science that focuses on the technical side of open government data. To identify open data policy attributes a socio-technical perspective is needed, thus the social science is chosen as the subject area over the more technical computer science area.

Other than that, the literature is also found through the snowballing method by looking at previous and subsequent study that cites the key literature such as Charalabidis, Alexopoulos, & Loukis' (2016) study about a taxonomy for OGD research.

Search terms:
(open AND government AND data AND (measurement OR assessment OR evaluation OR preference)) AND (LIMIT-TO (SRCTYPE , "j ") OR LIMIT-TO (SRCTYPE , "p ")) AND (LIMIT-TO (SUBJAREA , "SOCI ")) AND (LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013))

The initial search result in a total of 168 papers. A quick scan of the abstracts is performed on all papers to decide if they are relevant for this study. This step reduces the total number of papers to 44. Lastly, content analysis is performed, focusing on the aspect of the open data policy discussed and the perspectives of the study. This result in 13 papers reviewed in Table 5.

3.2. Open Government Data (OGD) research domain

The Open Government Data (OGD) research domain consists of a wide range of topics. A study by Charalabidis, Alexopoulos, & Loukis (2016) creates a taxonomy for OGD research and categorize the topics in four research area: 1) OGD Management and Policies, 2) OGD Infrastructures, 3) OGD Interoperability, and 4) OGD Usage and Value. The study of open data assessment can be classified into the OGD Usage and Value research area. This research area describes research topics of assessment studies from various perspectives such as OGD value and impact assessment, OGD readiness assessment, and OGD portals evaluation framework. The diverse OGD assessment research topics reflect two sides of OGD, the data provision by government and the data reusability by the citizens.

3.3. The tension between "stewardship" and "usefulness"

Government as the implementation agent of OGD program is faced with the inherent tension between the stewardship and usefulness principles in managing the information (Dawes, 2010). The stewardship principle focuses on the data provisioning dimension which addresses the issue of data confidentiality, information quality, information and system security, data management, and maintenance of data assets. On the other hand, the usefulness principle aims to foster the utilization of data to generate social and economic benefits which lead to strategies that improve public access to government information, stimulate public-private information partnerships and innovative application of data.

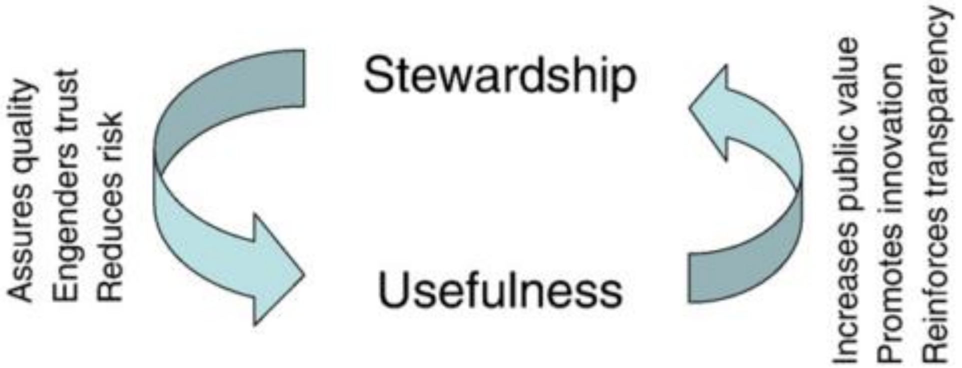


Figure 5 Conceptual model of information-based transparency principles adapted from (Dawes, 2010)

Both stewardship and usefulness principles are important to address the adoption barriers of open data. Janssen et al. (2012) categorize the barriers into the data provider barrier and data user barrier. Examples of data provider barriers are institutional barriers (risk-averse culture, limited resource to handle the data publishing process), whereas the data use barriers are task complexity (lack of metadata, lack of skills to discover the data) and use and participation (lack of incentives, insufficient knowledge to process the data). Stewardship principles to address the data provider barrier and usefulness principles to address the data user barrier.

These two principles may appear to be contradictory, but they are complementary. Although they serve different purposes, those purposes are compatible and can be mutually reinforcing. For example, data stewardship leads to better documentation and quality of the data provided which helps the potential users to find the datasets they need. The end users who create services using open data will also be critical on the quality, documentation, and up-to-date version of the datasets. However, given the limited (personnel and monetary) resource available for the implementation, policymakers and government agencies continuously face the challenge to balance these two principles.

Further study by Reggi & Ricci (2011) assess 434 beneficiaries of EU Structural Funds and found that the open data strategy of those beneficiaries diverges into two clusters resembling the tension between stewardship and usefulness principles. "User-centered" cluster focuses on the usefulness principle by providing data visualization and searching features, while "Re-user centered" cluster apply the stewardship principle by concentrating on data quality and validity. Lee & Kwak (2012) also differentiate data-related and participation/collaboration-related capabilities/processes in their 5 stage Open Government Maturity Model (OGMM), with the early stage focus on data capabilities and later stage on the participation/collaboration capabilities.

This tension is reflected in the existing benchmark and assessment studies where each study approach open data policy from diverse perspectives such as: the degree of dataset reusability, data quality, and other data provisioning attributes (Petychakis, Vasileiou, Georgis, Mouzakitis, & Psarras, 2014; Tim Berners-Lee, n.d.; Vetrò et al., 2016), open data portal content and features (Afful-Dadzie & Afful-Dadzie, 2017; Lourenço, 2015; Thorsby, Stowers, Wolslegel, & Tumbuan, 2017; Zuiderwijk-van Eijk & Janssen, 2015), user perspectives on open data usability (Weerakkody et al., 2017), and the holistic approach which assess the open data program as an ecosystem (Ubaldi, 2013; Welle Donker & van Loenen, 2017)

3.4. The different aspects and perspectives in open government data study

OGD initiatives need to address challenges from different aspects (policy, legal, economic, organizational, technical, and cultural) in order to create an ecosystem that enables value creation (Ubaldi, 2013). The multi-perspective nature of OGD initiatives is also reflected in aspects and perspectives investigated by different open government data studies.

Three main perspectives are identified: open data portal perspective, socio-technical perspective, and citizen perspective. The perspective can be described as a viewpoint that is taken by researchers in their study. For example, researchers can choose open data portals, the socio-technical institution in which the OGD is applied, or citizen perception of the OGD as the object of interest for their study.

In Table 5 the aspects and perspectives taken by each open government data study are summarized.

Open data portal perspective

Most of the studies use the open data portal perspective because it is the most common implementation of OGD initiatives. However, those studies can investigate different aspects of open government data even though they have the same object of interest (open data portal).

Sayogo, Pardo, & Cook (2014) analyze 35 countries open government data portals progress based its data manipulation and engagement capability. The progress status is based on the availability of features in the open data portal. For example, the portal is considered having advanced data manipulation capability if the portal provides tools that enable users to combine multiple datasets and do the data analysis in the portal; similarly, the portal obtains advanced engagement capability status if the portal provides features for inter users collaboration. The study concludes that OGD portals development follows an incremental approach similar to e-government development stage. The data manipulation and engagement capability are also found in the Open Government Maturity Model (OGMM) proposed by Lee & Kwak (2012).

Thorsby et al. (2017) compare 37 cities open data portal in America based on its features and content diversity. However, the definition of features in this study differs from Sayogo, Pardo, & Cook (2014). The study categorizes features into content, help, policy, and results; and each feature has a different category of measurement.

The content, help, and result feature are comparable to data manipulation and engagement capability from Sayogo, Pardo, & Cook (2014). Thorsby et al. (2017) divide the engagement capability into two features (help and result). The help features mainly assess the capability to search through datasets, tutorials, and contact information for help; the result features measure the portal ability to showcase the created application, promote and invite citizen to use open data, and the integration capability (API). The study also investigates the availability of open data policy and clear terms of use as the policy features.

Chatfield & Reddick (2017) examines 20 local governments open data portal in Australia based on its service capabilities. The service capabilities are as follow open data provision, data format variety, open data policy intensity, and entrepreneurial data services. The study shows that local government with medium and high open data policy intensity tend to have greater number of published data and provide services beyond the standard data provision. The entrepreneurial data service is described as active government involvement to foster citizen co-creation of open data services. For example, organizing hackathon events, providing data analytic tools, and users skills development (data analysis and data modeling).

On the other hand, some studies specifically investigate the characteristic of datasets provided in the open data portal. Lourenço (2015) measures seven national open data portal based on the desired characteristic of data it disclosed. The study measures the data quality, completeness, access and visibility, usability and comprehensibility, timeliness, value and usefulness, granularity, and comparability of the published datasets. Another study by Vetrò et al. (2016) analyzes the data quality of open data portal using the established data quality metrics such as completeness, accuracy, traceability, currentness, expiration, compliance, and understandability. Both studies investigate the datasets characteristic from the data provider side. Afful-Dadzie & Afful-Dadzie (2017) use data-related metrics (data quality, data format, metadata, data availability, data integrity) in 5 African countries open data portal to inquire journalists attributes preferences for the portal. They found that the respondents chose metadata as the most important attributes with the relative importance weight of 28.82%, followed by data format (23.3%) and data quality (20.34%).

Zuiderwijk & Janssen (2015) create a quasi-experiment to measure the effect of participation mechanism and data quality indicators in the open data portal, the participants are assigned into the control and treatment group. The control group is asked to use the prototype open data portal which includes the participation features such as discussion messages, social media sharing, linking items related to a dataset, wiki descriptions and discussions, and data quality ratings and reviews. The study suggested that participation mechanisms and quality indicators add value and improve the use of OGD portal.

The abovementioned studies analyze open data portal, and there are recurring aspects from those studies such as data manipulation capability, engagement capability, availability of open data policy, and non-technical features (promotion of open data, user's skills development, engagement events).

Socio-technical perspective

Next stream of study focusses on a more comprehensive approach of assessing OGD by considering the socio-technical aspects of OGD.

A study by Ubaldi (2013) provides an analytical framework and metrics of measurement on several dimension consist of policies and law, technical, data governance, organizational, communication and interaction, political priorities, impact, and data-related metric such as availability, quality, uptake, re-use. This framework is applied in a national scope and become the basis of the OECD survey on Open Government Data.

In the context of Dutch open data policy, two studies that apply socio-technical perspective are found. First, Zuiderwijk & Janssen (2014) creates a framework for comparing OGD implementation in seven Dutch government organizations. The study analyzes several aspects such as policy environment and context, policy content, performance indicators, and public values. The analysis is conducted from the data provider side (government) which can be seen from the information that is being measured by each aspect. Examples of information being compared for the policy environment and context: level of government organization, resource allocation, legislation, socio-political context, the culture of the institutions. The policy content aspect provides the specification of the OGD such as target groups of open data, policy strategy and principles on publishing data, technical standards and formats of open data. The study found that most of the policies investigated focus on internal challenges to publish the data (privacy protection, confidentiality, data misuse and misinterpretation, embargo periods, data quality, data completeness) and less concern on the usability of the data (how it can be used to create the desired public values.

Second, Welle Donker & van Loenen (2017) examines the Dutch open data ecosystems from two aspects: data supply indicators (known, attainable, usable) and data governance indicators (vision, leadership, self-organizing ability, financing, open data stimulation, supply-user communication, G2G communication). The data supply indicators investigate whether the dataset is searchable and can be found for use (known), accessible from a financial, legal, and practical aspect (attainable), and (usable) in terms of having complete metadata, documentation, and up-to-date. The study not only analyzes open government data from the government perspective but also ask the infomediaries (users who developed services using open data) about the open data governance. The study found that infomediaries criticize the existing data governance model where government is waiting for the creation of “killer app” and organize hackathon with temporarily available datasets. The infomediaries would prefer the government to develop a sustainable open data business model; being a launching customer and commission the infomediaries to develop open data tools and applications.

Dawes, Vidiasova, & Parkhimovich (2016) assess OGD programs in New York and St. Petersburg within the dimensions of settings, motivation, policy and strategy, data publication and use, feedback and communication, benefits, and advocacy and interaction among stakeholders. The dimensions used are comparable with the abovementioned study; for example, settings, motivation, and policy & strategy are similar to Zuiderwijk & Janssen (2014) policy environment and context and policy content. The data publication and use are comparable with data supply indicators (known, attainable, usable) from Welle Donker & van Loenen (2017) study.

Studies in the socio-technical streams mainly focus on the policy context and strategy of the government. However, it also discusses the data-related attributes and participation & engagement related attributes extensively. Even though it uses different terminology, for example, data supply/policy content/data publication and use for the data-related attributes; communication and interaction/feedback and communication/open data stimulation/supply-user communication for the participation & engagement related attributes.

Citizen perspective

Recent studies start to investigate open government data from the citizen perspectives. Weerakkody et al., (2017) measures the citizen intention to use open data using the modified and extended Diffusion of Innovation (DOI) model with predictors: relative advantage, compatibility, observability, and security risk. The study found that relative advantage, compatibility, and observability are statistically significant in predicting citizens intention to use open data. The security risk had no significant effect on citizen intention to use open data. The study suggests that most citizens have no concerns about trusting public sector open data and do not perceive a significant security risk in the open data.

Zuiderwijk et al. (2018) investigate the attainment of OGD objectives based on the delivered benefits which are categorized into operational, technical, economic, and societal benefits. The study shows that the most delivered benefits are operational and technical benefits, followed by economic benefits, and societal benefits. The study also concludes that there is a mismatch between open data objectives and the delivered benefits. Achievement of the benefits is not significantly related to the presence of objective related to the delivery of the benefits.

Safarov, Meijer, & Grimmelikhuijsen (2017) conduct a systematic literature review on the utilization of open government data and identify the conditions for utilization. In the study, they review 101 studies and found two categories of condition for utilization which are a technical and social condition. Technical conditions refer to the feature of OGD such as the data quality, data availability, and infrastructure to enable OGD; Social conditions refer to the institutional context (policy, legislation, organization) and the skills of users. The study also found the distinction between users, direct users who use the OGD themselves and indirect users who use the data/services processed by intermediaries.

Table 5 Summary of assessment studies

Study	Method	Object of Assessment	Measured aspects	Perspectives
(Afful-Dadzie & Afful-Dadzie, 2017)	Quantitative (survey)	Open data portal	Journalist preferences of data-related metrics: • data quality • data format • metadata • data availability • data integrity	Citizen (journalist)
(Chatfield & Reddick, 2017)	Quantitative	Open data portal	• open data provision • data format variety • open data policy intensity	Government

			• entrepreneurial data services.	
(Thorsby et al., 2017)	Quantitative (scoring)	Open data portal	Open data portal features and content diversity. Features category: • content • help • policy • results	Government
(Welle Donker & van Loenen, 2017)	Qualitative	Holistic (data supply, data governance, user)	Data supply indicators: • Known • Attainable • Usable Data governance indicators: • Vision • Leadership • self-organizing ability • financing • open data stimulation • supply-user communication • G2G communication	Government and Citizen
(Lourenço, 2015)	Qualitative	Open data portal	Data disclosure characteristics: • quality • completeness • access and visibility • usability and comprehensibility • timeliness • value and usefulness • granularity • comparability	Government
(Zuiderwijk et al., 2018)	Quantitative (survey)	Relation of OGD initiatives and delivered benefits	Four categories of delivered benefits: • Operational • Technical • Economic • Societal	Citizen
(Safarov et al., 2017)	Qualitative (Literature Review)	Discussion of open data utilization in the academic community.	conditions for utilization: • quality of data • legislation/policy • skills • infrastructure • availability • privacy	Academic

(Zuiderwijk-van Eijk & Janssen, 2015)	Quasi-Experiment	Open data portal	participation mechanism and data quality indicators: • discussion messages • social media sharing • submissions of related items • wiki descriptions and discussions • data quality ratings • data quality reviews	Citizen
(Vetrò et al., 2016)	Quantitative	Open data portal	Data quality: • Completeness • Accuracy • Traceability • Currentness • Expiration • Compliance • Understandability	Government
(Weerakkody et al., 2017)	Quantitative (survey)	Citizen intention to use open data	• relative advantage • compatibility • observability • security risk	Citizen
(Ubaldi, 2013)	Qualitative	Holistic	• policies and law • technical • data governance • organizational • communication and interaction • political priorities • impact • data-related metric such as availability, quality, uptake, re-use.	Government
(Sayogo et al., 2014)	Quantitative	Open data portal	• data content • data manipulation capability • participatory and engagement capability	Government
(Dawes et al., 2016)	Qualitative	Holistic	• policy and strategy • data publication and use • feedback and communication • benefit generation • advocacy and interaction among stakeholders	Government

3.5. Identification of potential trade-off attributes

In Table 6, the identified assessment attributes from existing studies are presented. Reviewed studies in section The different aspects and perspectives in open government data studyrepeatedly use the variance of data-related and participation & engagement related aspects in their analysis. Therefore, in this study, the trade-off attributes are categorized into data-related attributes, and participation & engagement attributes as shown in Table 6**.** The categories also reflect the tension between data stewardship and usefulness principles discussed in section 3.3. Other than that, there are attributes specifically related to the usability, communication, and interaction features of the open data portal.

In the identification process, this research only selects attributes which can be experienced directly by the citizens. Therefore, aspects that discuss the internal arrangement of the data provider are excluded. For example, intergovernmental agency communication, organization restructuring, political priority.

Table 6 Open Data Policy attributes from existing assessment study

Category	Attributes	Study
Data-related attributes	Data Availability (number of datasets, API)	(Afful-Dadzie & Afful-Dadzie, 2017; Petychakis et al., 2014; Safarov et al., 2017; Sayogo et al., 2014; Thorsby et al., 2017; Ubaldi, 2013; Welle Donker & van Loenen, 2017)
	Data Quality (accuracy, consistency, update timeliness, completeness)	(Afful-Dadzie & Afful-Dadzie, 2017; Petychakis et al., 2014; Safarov et al., 2017; Thorsby et al., 2017; Ubaldi, 2013; Vetrò et al., 2016; Welle Donker & van Loenen, 2017; Zuiderwijk-van Eijk & Janssen, 2015)
	Data Discoverability (advanced search tools on portal, metadata)	(Afful-Dadzie & Afful-Dadzie, 2017; Attard, Orlandi, Scerri, & Auer, 2015; Petychakis et al., 2014; Thorsby et al., 2017; Welle Donker & van Loenen, 2017)
	Data Protection	(Attard et al., 2015; Safarov et al., 2017; Weerakkody et al., 2017)
Portal-related attributes	Communication and Interaction in Open Data Portal	(Petychakis et al., 2014; Safarov et al., 2017; Sayogo et al., 2014; Thorsby et al., 2017; Titah, 2017; Ubaldi, 2013; Zuiderwijk-van Eijk & Janssen, 2015)
	Open Data Portal ease of use	(Safarov et al., 2017; Thorsby et al., 2017; Titah, 2017; Weerakkody et al., 2017)

Participation & engagement related attributes	Public Awareness	(Attard et al., 2015; Thorsby et al., 2017; Weerakkody et al., 2017; Welle Donker & van Loenen, 2017)
	Public Participation (citizen involvement in promoting, using, and discussion about open data)	(Attard et al., 2015; Titah, 2017; Welle Donker & van Loenen, 2017)
	Motivation (competition, public-private partnership)	(Attard et al., 2015; Weerakkody et al., 2017; Welle Donker & van Loenen, 2017)
	Development of required skills and expertise to use Open Data	(Safarov et al., 2017; Welle Donker & van Loenen, 2017)
	Compatibility (the provided open data suit the needs of the citizen)	(Weerakkody et al., 2017; Welle Donker & van Loenen, 2017)
	Data Reusability (number of applications created, number of new services from open data)	(Sayogo et al., 2014; Thorsby et al., 2017; Ubaldi, 2013)

3.6. Conclusion

The literature review identifies a tension between 'stewardship' and 'usefulness' principles in the open data policy. This tension is mapped as data-related and participation/collaboration-related capabilities/processes in Open Government Maturity Model (OGMM) by Lee & Kwak (2012). Furthermore, the existing open data policy assessment study approaches the open data policy from diverse perspectives which are open data portal perspective, the socio-technical perspective, and citizen perspective.

Answering the sub research question posed at the beginning of this chapter: *What are the possible trade-off attributes for the open data policy in the existing literature?*

Three common categories of open data policy attributes are identified from the literature review: data-related attributes, portal-related attributes, and participation & engagement related attributes.

- **Data-related attributes** consist of data availability, data quality, data discoverability, and data protection.
- **Portal-related attributes** are communication and interaction features, open data portal ease of use.
- **Participation & engagement related attributes** are public awareness, public participation, motivation, development of required skills and expertise, compatibility of the data provided with the needs, and data reusability.

The attributes are selected based on its possibility to be directly experienced by the citizens. If the citizens have experience related to the attributes it will help them to understand the survey and give a valid response. Therefore, attributes that are hardly perceived by the citizens and related to the internal arrangements of data providers are excluded. For example, vision and leadership, organization restructuring, interagency communication, legislation.

In the next chapter, this category of attributes will be explored in combination with the context of open education data in the Netherlands from Chapter 1. The suitable attributes will be selected for the design of citizen stated choice experiment.

Chapter 4: Citizen stated choice experiment design

This chapter is aimed to answer sub research question 3: *How do the identified trade-off attributes and policy context translate into the citizen stated choice experiment design?*

In this chapter, the process of designing the citizen stated choice experiment is discussed. It starts with the explanation of the attribute selection process, based on the information collected from Chapter 1 (Dutch open education data policy context) and Chapter 3 (potential trade-off attributes from the literature). Next, the attribute levels of the selected attributes are specified, these attribute levels become the basis for the policy alternatives formulated in the survey design. Afterward, the process of creating the survey and conducting pilot test are discussed. Finally, the specification of the final survey is explained.

4.1. Attribute selection

The selection of attributes is based on several criteria:

- Expected influence on an individual choice (in this context the Dutch higher education students as the target respondents)
- Societal relevance of the factor (whether the attributes complement the Dutch open education data policy motivation for *education quality openness*)
- Measurability in the discrete choice experiment (whether the attributes have a tangible unit of measurement and can be operationalized for the choice situations)

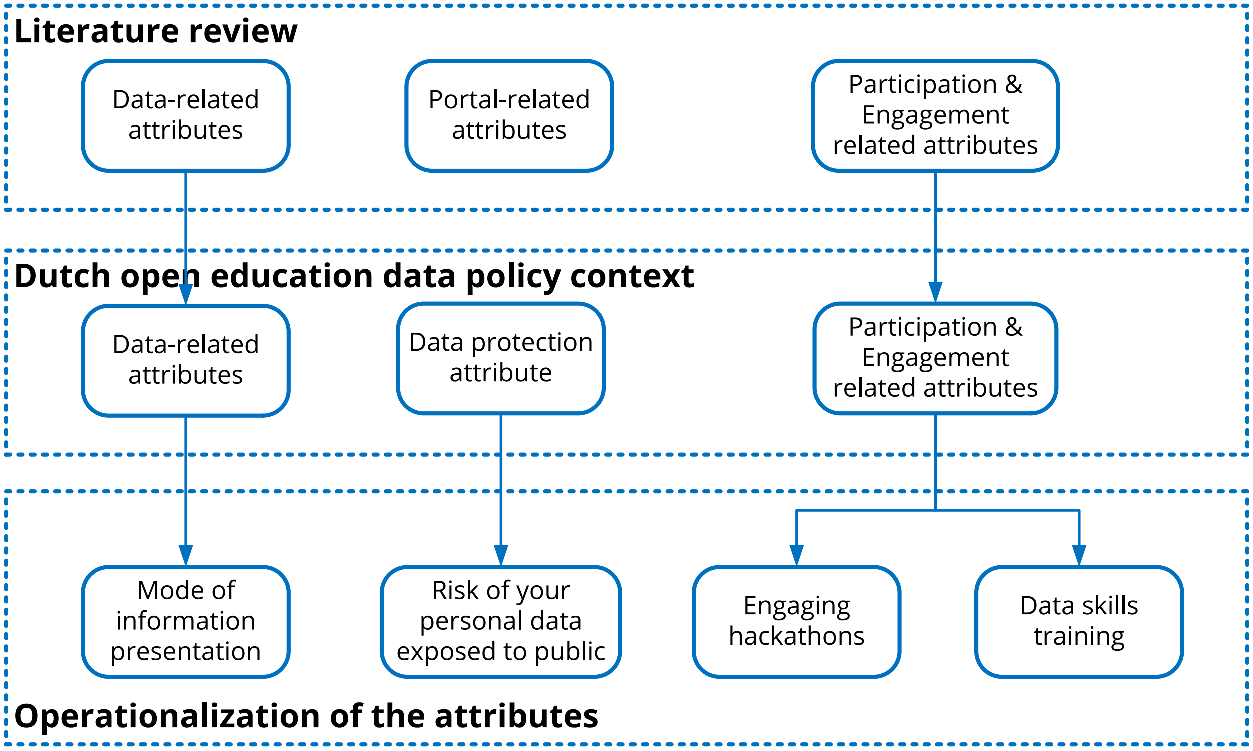

Literature review
Data-related attributes
Portal-related attributes
Participation & Engagement related attributes
Dutch open education data policy context
Data-related attributes
Data protection attribute
Participation & Engagement related attributes
Mode of information presentation
Risk of your personal data exposed to public
Engaging hackathons
Data skills training
Operationalization of the attributes

Figure 6 Conceptual framework of open education data policy attributes

Figure 6 shows the conceptual framework of open data policy attributes. The attributes selection is based on the combination of potential trade-off attributes in Chapter 3 and the Dutch open education data policy context in Chapter 1.

Categories of open data policy attributes

Three categories of potential open data policy attributes are identified in Chapter 3 which are *data-related attributes*, *portal-related attributes*, and *participation & engagement attributes*. Based on the policy context exploration, there are significant implementation of data-related attributes and participation & engagement attributes within Dutch open education data policy.

The Ministry of OCW provides information in diverse forms such as raw data in the respective open data portals (DUO, OCW, Education Inspection Agency), static and interactive figures (OCW and VSNU portals), and creating services from open education data (scholenopdekaart.nl and studiekeuze123.nl). Other than that, several participation & engagement events are organized such as data exploration event “Education Data under scrutiny” and hackathon “Hack de Valse Start”. There is no specific portal-related attributes implementation in the OCW open education data policy; all the data are simply hosted in each agency open data portal without any additional features for the users to interact with the portal.

However, one aspect of data-related attributes is growing in importance based on the policy context. The increasing importance of data protection attribute is influenced by the passing of the General Data Protection Regulation (GDPR) on 25th May of 2018.

The introduction of GDPR reinforces the existing barrier faced by government agencies in opening their data (risk-averse culture and limited resource to handle the data publishing process). The risk of opening data is increased due to a hefty fine in case of data breaches. Sizable resources are required (both human resources and monetary) to fulfill the GDPR data protection specification. This condition put pressure on the already limited budget and personnel of government agencies in charge of open education data (DUO, OCW, Education Inspection Agency). The complexity of opening data is increased due to the additional requirement of Data Protection Impact Assessments (DPIA).

Based on the policy context exploration the three categories of open education policy attributes is modified into data-related attributes, data protection attribute, and participation & engagement related attributes. The portal-related attributes are omitted because in the context of open education data policy there is only a basic open data portal implementation.

Operationalization of the selected attributes

Since the three identified categories are still abstract, this section discusses the operationalization of those categories into tangible attributes for the survey design.

For the data-related category, mode of information presentation is selected as the attribute. It is assumed that the provided data meets the data quality standard (accuracy, consistency, update timeliness, completeness), complete metadata, and accessible in a standard format.

The participation & engagement related attribute is the umbrella term for a diverse type of activities to stimulate the public participation such as public training to increase the citizen data proficiency, a hackathon to create new services, data exploration event to identify public data needs, support for a monthly meeting of civic innovators. Therefore, two attributes are defined for the participation & engagement related attribute which are: engaging hackathons and data skills training

Finally, the risk of respondent personal data exposed to the public is selected for the data protection category

The discrete choice experiment considers labeled and unlabelled alternatives. Labeled alternatives are used when the labels represent characteristics not varied in the experiment. For example, DCE for the mode of transportation has label specific characteristic such as car, train, plane. Each alternative has specific characteristics that are not varied, or there are alternative specific attributes, e.g., different range for travel time, parking fee for the car.

In this experiment the unlabelled alternatives are used because both alternative use the same generic attributes and there is no label specific characteristic, the alternatives are simply called Policy A and Policy B.

Table 7 Overview of the level of measurement and unit of measure

Category	Level of Measurement	Attributes
Data-related attribute	Nominal	Mode of information presentation
Participation & engagement related attribute	Ratio	Number of free engaging hackathon events Number of free citizen data skill training events
Data protection attribute	Ratio	Risk of your personal data exposed to the public

Table 8 Specification of parameters for pilot design

Attributes	Parameter
Mode of information presentation	β_1Data
Engaging hackathons	β_2Hackathon
Data skills training	β_3Training
Risk of your personal data exposed to public	β_4Privacy

4.2. Attribute levels

Data-related attribute

Access to the information attribute reflects the different mode of information presentation that is currently implemented. In the DUO portal, there are ten published data entries for Higher Education which consists of an address, a number of registered students, and the financial details of Higher Education institutions. The Education Inspection Agency publish nine datasets (5 quality indicators, two final assessment of the institution, one list of excellent schools, and one list of very weak schools). Other than that, OCW publishes static and dynamic figures in its websites (onderwijsincijfers.nl, ocwincijfers.nl, trendsinbeeldocw.nl). In the survey, three different mode of information presentation are available: in original form (as similar as possible to the source) as the base value, static or interactive figures, and functional services (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl).

Participation & engagement related attribute

OCW have an annual event for education knowledge sharing called Kennismarkt (Knowledge Market) where the participants (policymakers, practitioners, and researcher) can attend the lecture, workshop, and discussion regarding the future of education program. Other than that, OCW also arranges several participation and engagement events. In 2016, OCW organized "Education Data under scrutiny" which brings parents, students, teachers, and school management together to discuss the possible application of open data (Rijksoverheid, 2016). In 2018, OCW and municipality of Amsterdam organize a hackathon, "Hack de Valse Start", to gain insight on unequal opportunities by combining education and municipality open data provided by DUO and CBS (openstate.eu, 2018).

In the survey, the participation & engagement events are represented by two attributes: free engaging hackathon event and free citizen data skill training event. The reason behind it is to provide more concrete attributes for the respondent to compare rather than a generic term of participation & engagement events. The term hackathon and data skill training can be specified in its aim and the benefits provided.

Therefore, the experiment chooses 1 free engaging hackathon event per 2 years and 1 free citizen data skill training event per year as the base value for participation & engagement events attribute. The attribute levels are scaled up to (1 event per year and 2 events per year) for free engaging hackathon events and (2 events per year and 3 events per year) for free citizen data skill training events.

Data protection attribute

OCW annual report in 2017 record that there are 47 cases of data breaches reported within DUO and 3 cases of data breaches in OCW (OCW, 2018). 20 cases of the data breaches in DUO have been reported to the Dutch Data Protection Authority according to the regulations. There is no information about when the data breach happens, what type of data is compromised, and from what channel the data breach happens.

The existing open education data is highly deanonymized and only contain the aggregate information which cannot be traced to the individual. However, if there is a need for fine-grained data for a particular use case such as a hackathon that requires the social background information of the students, the data will be more susceptible to be compromised. Therefore, the experiment chooses 1 incident per year as the base value followed by 1 incident per quarter and 1 incident per month as the range for the number of data leak incidents.

Table 9 Overview of attribute level and value

Category	Attributes	Value
Data-related attribute	Mode of information presentation	• in original form (as similar as possible to the source)

		• as static or interactive figures • as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)
Participation & engagement related attribute	Number of free engaging hackathon events	• 1 every 2-years • 1 per year • 2 per year
	Number of free citizen data skill training events	• 1 per year • 2 per year • 3 per year
Data protection attribute	risk of your personal education data exposed to the public	• 1 incident per year • 1 incident every 3-months • 1 incident per month

4.3. Pilot Survey

The pilot survey is the first phase of the survey design process to test the survey with 10 respondents and collect feedback on the survey length and understandability. The feedback from respondents can be used to adjust attribute levels. The following paragraphs elaborate on the different design steps to design a pilot survey. These steps are:

1. Model specification
2. Generating experimental design
3. Constructing the survey

Model Specification

The first step in the design of a stated choice experiment is the specification of the model. The pilot study model contains two unlabelled alternatives, labeled attributes and no alternative specific constant (ASC). The utility functions for the two alternatives are shown in the equation:

$$U(alt1, alt2) = B\_Wdata * Xdata + B\_Whackathon * Xhackathon + B\_Wtraining * Xtraining + B\_Wprivacy * Xprivacy + \varepsilon$$

Variable	**Definition**
U(alt1, alt2)	Utility function for policy A and policy B
B_Wdata	Generic parameter for the attribute mode of information presentation
B_Whackathon	Generic parameter for the attribute free engaging hackathon events
B_Wtraining	Generic parameter for the attribute free engaging data skill training events

B_Wprivacy	Generic parameter for the attribute data protection
ε	Random error component

Generation of experimental design

The second step in the survey design of the pilot study is the generation of the experimental design. This design shows the set of combination of attribute levels that respondents base their hypothetical choice on. There are several types of experimental designs: orthogonal designs, efficient designs and Bayesian designs. Efficient and Bayesian designs require prior information on the utility coefficients of the different parameters. An orthogonal design assumes that attribute levels are not correlated and therefore sets prior values to zero. No previous literature assesses the citizen preferences for open data policy. Therefore, an orthogonal experimental design is generated.

A fractional factorial orthogonal design is selected to estimate the most reliable parameters with the lowest standard errors. Full factorial designs are not feasible because this leads to too many choice situations: 3^4 = 81. Furthermore, the research budget does not allow to block the experiment. Therefore, a basic plan 2 design is chosen, with three attributes in three levels and a total of 9 choice sets. The advantage of this plan is its simple orthogonal (reliable) design that measures all main effects and maintains attribute level balance.

Sequential construction of the alternatives and the choice sets is used since there is no alternative specific attribute. The software package Ngene is used to generate the orthogonal design for the pilot study. As discussed, this is a fractional factorial design (basic plan 2) with four attributes, each attribute has three attribute levels, and no attribute specific constant (ASC). Table 10 shows the overview of the nine choice sets.

Table 10 Overview of choice situations

Design	Alternative 1				Alternative 2			
Choice situation	Mode of information presentation	Number of free engaging hackathon events	Number of free citizen data skill training events	risk of your personal education data exposed to the public	Mode of information presentation	Number of free engaging hackathon events	Number of free citizen data skill training* events	risk of your personal education data exposed to the public
1	in original form (as similar as possible to the source)	1 every 2-years	1 per year	1 incident per year	as static or interactive figures	1 per year	1 per year	1 incident every 3-months
2	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	1 per year	2 per year	1 incident per year	as static or interactive figures	2 per year	3 per year	1 incident per year

3	as static or interactive figures	2 per year	3 per year	1 incident per year	in original form (as similar as possible to the source)	1 per year	3 per year	1 incident per month
4	as static or interactive figures	1 per year	1 per year	1 incident every 3-months	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	1 every 2-years	3 per year	1 incident every 3-months
5	in original form (as similar as possible to the source)	2 per year	2 per year	1 incident every 3-months	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	1 per year	2 per year	1 incident per year
6	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	1 every 2-years	3 per year	1 incident every 3-months	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	2 per year	1 per year	1 incident per month
7	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	2 per year	1 per year	1 incident per month	as static or interactive figures	1 every 2-years	2 per year	1 incident per month
8	as static or interactive figures	1 every 2-years	2 per year	1 incident per month	in original form (as similar as possible to the source)	1 every 2-years	1 per year	1 incident per year
9	in original form (as similar as possible to the source)	1 per year	3 per year	1 incident per month	in original form (as similar as possible to the source)	2 per year	2 per year	1 incident every 3-months

Constructing the survey

The generated experimental design shows the combination of attribute levels that are presented to respondents. Every row from Table 10 is transformed into a choice situation in the survey. The online survey program SurveyGizmo is used to design the full pilot survey. The survey is constructed in English and distributed to Dutch higher education students within the network of friends.

The pilot survey consists of three different parts:

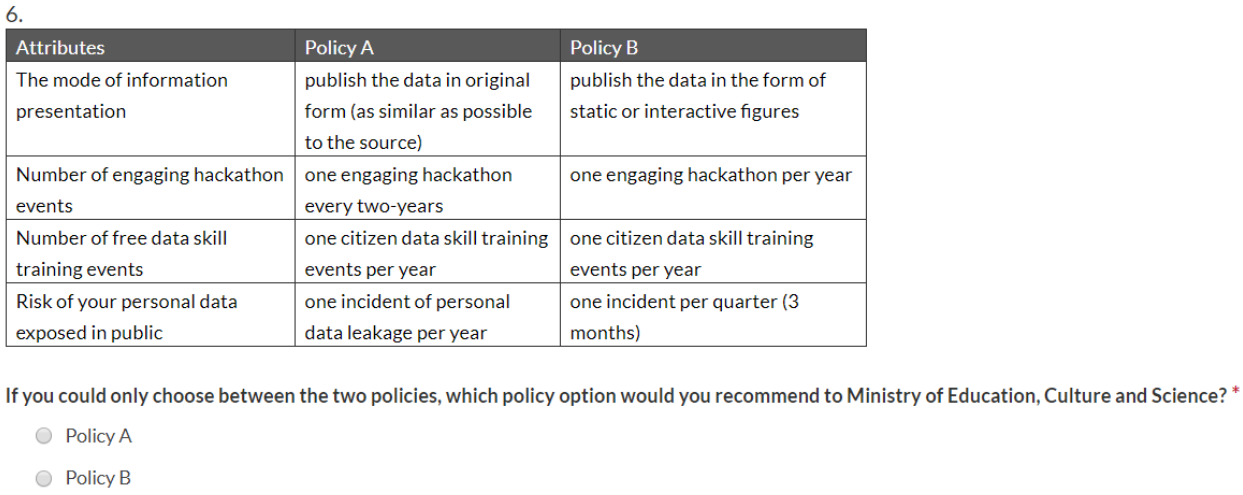
1. Leading questions about open education data policy
   The leading questions are aimed to guide the respondents through the attributes of open education data policy that they will compare in the choice situations. Alternatively, the description of attributes can be put in the form of long introduction paragraph however it might result in respondents skipping the description altogether. Therefore, leading questions are used to conceal the context introduction in a gradual approach that cannot be skipped. The questions result in information about the respondent familiarity with open education data attributes. The leading questions are presented in a neutral wording and consist of all the attributes in the choice situations to avoid "anchoring effect" where the respondents rely too heavily on an initial piece of information in their subsequent judgment. Furthermore, the questions are formulated

in an inquisitive manner about the respondent's experience rather than providing descriptions that could lead the respondent to favor one attribute over the others.

The leading questions are as follow:

- Have you ever searched for open education data?
- Which portals providing open education data have you ever accessed? (multiple options)
- Which services created using open education data have you ever used? (multiple options)
- Which open data events organized by the government have you ever participated in? (multiple options)
- On a scale from 1 to 7, to what extent are you concerned that the government will violate your privacy through the leakage of your personal data?

2. Choice situations
The main part of the survey where respondents choose between the alternatives based on the attributes of open education data policy. Figure 7 shows an example of the pilot study choice situation.



6.

Attributes	Policy A	Policy B
The mode of information presentation	publish the data in original form (as similar as possible to the source)	publish the data in the form of static or interactive figures
Number of engaging hackathon events	one engaging hackathon every two-years	one engaging hackathon per year
Number of free data skill training events	one citizen data skill training events per year	one citizen data skill training events per year
Risk of your personal data exposed in public	one incident of personal data leakage per year	one incident per quarter (3 months)

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B

Figure 7 Example of pilot study choice situation

3. Perception and demographic questions
The next part of the survey consists of questions that measure the respondent perception towards the attributes provided in the choice situations and the whole survey. The perception and demographic questions are used to collect the following information:

- the most and least important attribute for the respondent
- the difficulty, realism, and relevance of the survey
- Generate a feedback report to evaluate and improve the survey for the final survey design
- Demographic questions: age, gender, level of education and specialization

Results of the pilot survey

The pilot survey is distributed to selected respondents and collects responses from 10 respondents, seven male and three female respondents all of them from Complex Systems Engineering and

Management program of the Delft University of Technology with different specializations in Transport, Energy, Building & Spatial, and ICT.

The summary of feedback from the respondents (See Appendix A. Pilot test feedback for the complete feedbacks):

1. Respondents from non-ICT related background select "risk of your personal data exposed to the public" as the most important attributes. Majority of the respondents have questions about the details of hackathon and data skills training (what is the purpose of the events, how it will be conducted, what is the tangible benefit)
2. Reduce the wordiness of attribute levels and change the number from words to real numbers (instead of one, two, three → 1,2,3)
3. Define the extent of personal data leakage in the open education data (bank account data leakage will have a different impact than the possibility of identifying a person by combining multiple anonymous data)

4.4. Final Survey

The final survey design is based on the improvements suggested by the respondents and the respondent's choice behavior. In this part, the improvements based on the feedback are discussed.

Including the description for each attribute

In the pilot test survey, before asking the respondent to answer the choice situations, the overview of attributes is presented to them. Based on the feedback, respondents ask for a better attributes explanation hence the overview is modified by including a description of each attribute and its effect as shown in Table 11

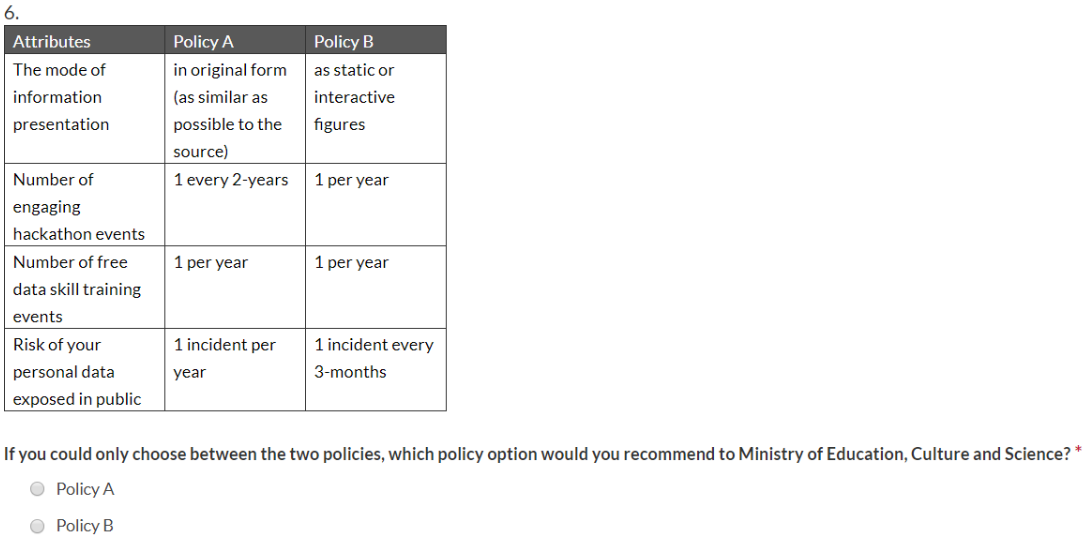
Table 11 Overview of attributes (Final survey)

#	Attributes	Options	Description
1	mode of information presentation	• in original form (as similar as possible to the source) • as static or interactive figures • as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)	• Data in original form is easier to be transformed into different forms (figures, input for other services) but harder to interpret and needs to be processed before it can be used. • Static or interactive figures are easier to interpret but harder to be transformed into different forms. • A service is an application created for a specific purpose. For example, studiekeuze123.nl to help students choose suitable study programs, scholenopdekaart.nl to help parents choose a primary and secondary school for their children
2	Number of free engaging hackathon events	• 1 every 2-years • 1 per year • 2 per year	The hackathon is organized by the government to address a specific social problem using the open education data. The results can be recommendations for the government or a prototype of service to address the problem.

			For example, Hack de Valse Start hackathon aimed to gain more insight with the help of data on how municipalities and school boards can identify and tackle inequality of opportunity in education.
3	Number of free citizen data skill training events	• 1 per year • 2 per year • 3 per year	basic training to improve citizen data literacy (ability to understand, use and communicate data effectively). Examples of data skills: • searching for the data • combining one dataset with other datasets • data interpretation • identify potential services that can be created from the datasets • identify potential datasets that have not been published yet
4	risk of your personal education data exposed to the public	• 1 incident per year • 1 incident every 3-months • 1 incident per month	The open education data is anonymized. The personal data leakage happens when a person can be identified by the combination of multiple anonymous open datasets.

Reducing the wordiness of choice situation

In the pilot survey, the respondents suggest reducing the wordiness of attribute value in the choice situation for better understandability. Figure 8 shows the choice situation in the final survey.



6.

Attributes	Policy A	Policy B
The mode of information presentation	in original form (as similar as possible to the source)	as static or interactive figures
Number of engaging hackathon events	1 every 2-years	1 per year
Number of free data skill training events	1 per year	1 per year
Risk of your personal data exposed in public	1 incident per year	1 incident every 3-months

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

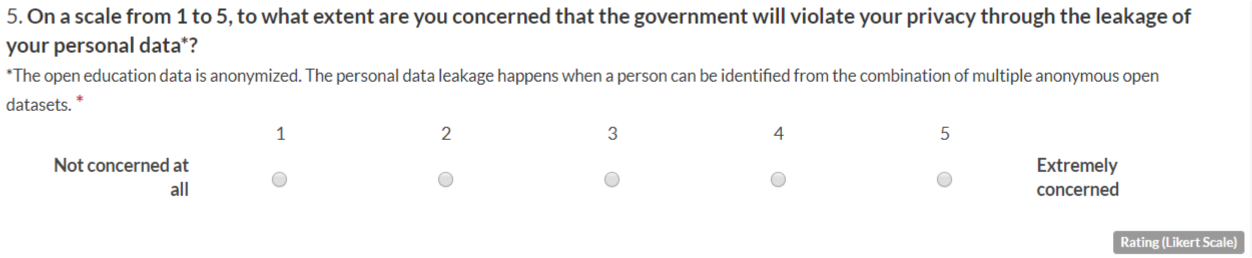
○ Policy A

○ Policy B

Figure 8 Example of the choice situation in the final survey

Clearly define the extent of data leakage

In the pilot survey, the respondents also asked about the extent of data leakage to understand its impact on their privacy. Therefore, the leading question for the data leakage attribute is modified to include the description of open education data leakage as shown in Figure 9. The Likert scale is modified into 1 to 5 scale since it is sufficient to capture the respondent perception toward the risk of data leakage.



5. **On a scale from 1 to 5, to what extent are you concerned that the government will violate your privacy through the leakage of your personal data*?**

*The open education data is anonymized. The personal data leakage happens when a person can be identified from the combination of multiple anonymous open datasets. *

	1	2	3	4	5	
Not concerned at all	○	○	○	○	○	Extremely concerned

Figure 9 Data leakage leading question

4.5. Conclusion

Answering the sub research question posed at the beginning of this chapter: *How do the identified trade-off attributes and policy context translate into the citizen stated choice experiment design?*

The citizen stated choice experiment is designed based on three categories of attributes identified in Chapter 3: data-related attributes, portal-related attribute, and participation & engagement attributes. Based on the policy context identified in Chapter 1 the portal-related attribute is omitted because there is the limited implementation of open education data portal. The open education data is simply hosted in the respective government agency portal (DUO, Education Inspection Agency, OCW) without any features for user interaction (visualization, data analysis).

However, there is an increasing importance for one of the data-related attributes which are the data protection. It emerges as a significant attribute due to the passing of the General Data Protection Regulation (GDPR) in 25th May of 2018. Government agencies face increasing barriers (risk and limited resource) in opening data.

The final selection of attributes are data-related attributes, data protection attribute, and participation & engagement attributes. Each of the attributes is further specified into measurable options, *"mode of information presentation"* for the data-related attributes, *"risk of your personal education data exposed to the public"* for the data protection attribute, and *"Number of free engaging hackathon events"* and *"Number of free citizen data skill training events"* for the participation & engagement attributes.

These attributes are then used to generate a fractional factorial orthogonal design with nine choice situations. Basic plan 2 design is chosen, with three attributes in three levels and a total of 9 choice sets. The experiment is generated using Ngene software with a sequential construction of the alternatives.

After that, the pilot survey consists of three parts are constructed: 1) Leading questions about open education data policy, 2) Choice situations, and 3) Perception and demographic questions. The pilot survey is tested among ten respondents, and the feedbacks are incorporated in the final survey.

The final survey is improved based on the following feedbacks: 1) include the description for each attribute, 2) reduce the wordiness of choice situations, and 3) clearly define the extent of data leakage.

The final survey from this chapter is distributed to the respondents. In the next chapter, the survey sampling procedure and the result of citizen stated choice experiment is discussed.

Chapter 5: Citizens preferences for a Dutch open education data policy

This chapter is aimed to answer sub research question 4: *What is the valuation of each trade-off attributes for the respondents in their role as a citizen?*

In this chapter, the sampling procedure of the survey distribution and the demographic profile of the respondents are explained. Next, the descriptive results of the citizen stated choice experiment is discussed. The descriptive results consist of several parts such as respondent familiarity with open education data, the most and least important attributes, choice distributions, and respondents' perception toward the experiment.

After that, the collected observations are used to generate the Multinomial Logit (MNL) model and infer the valuation of each trade-off attributes for the respondents. Finally, the qualitative results (feedback/comments) from the respondents are discussed.

5.1. Sampling procedure

The final survey is distributed among students who are currently attending a Dutch higher education institution or recently graduated. The higher education students are targeted due to several reasons:

- Higher education students have a relevant use case for the open education data which make them more likely to know about open data. (e.g., use open education data for courses, use the service to search for study programme)
- Higher education students have relevant skills to use open education data which make them more likely to be motivated by using open education data. (e.g., data analysis skill, programming skill)
- Higher education students are more likely to understand the term used in the survey with a proper explanation compared to other potential respondents (i.e., parents, primary/secondary school students).

The online survey is distributed through the network of friends and self-distributed in the Delft University of Technology. The survey obtained 59 respondents from 18-30 June 2018. The summary of sample characteristics is presented in Table 12.

Table 12 Sample characteristics

Gender	Count	Percentage	
Male	47	79.66%	
Female	11	18.64%	
I do not want to specify	1	1.69%	
Age	Count	Percentage	
18 - 24	39	66.10%	
25 - 30	18	30.51%	
Above 30	2	3.39%	
Education	Count	Percentage	Specialization

HBO (hoger beroepsonderwijs)	6	10.17%	I do not want to specify = 6 Business & Economics = 1 Law = 1 Building engineering = 1 Educational studies = 1 Information Science = 1
WO (wetenschappelijk onderwijs)	53	89.83%	I do not want to specify = 6 Electrical engineering = 2 Engineering and Policy Analysis = 3 Complex Systems Engineering and Management = 9 Complex Systems Engineering and Management (ICT) = 2 Complex Systems Engineering and Management (B&S) = 1 Complex Systems Engineering and Management (Energy) = 2 Complex Systems Engineering and Management (T & L) = 1 Architecture = 1 Economics = 1 Civil Engineering = 4 Industrial Engineering and Management (IEM) = 1 Mechanical Engineering = 6 Technology, Policy, and Management = 5 Chemical Engineering = 1 System and Control = 1 Clinical Technology = 2 Computer Science = 2 Microbiology = 1 Economics = 1 Design for Interaction = 1

5.2. Descriptive results

Respondents familiarity to open education data

From the leading questions section before the choice situations, information about respondents familiarity with the open education data attributes is collected as shown in Figure 10 and Figure 11.

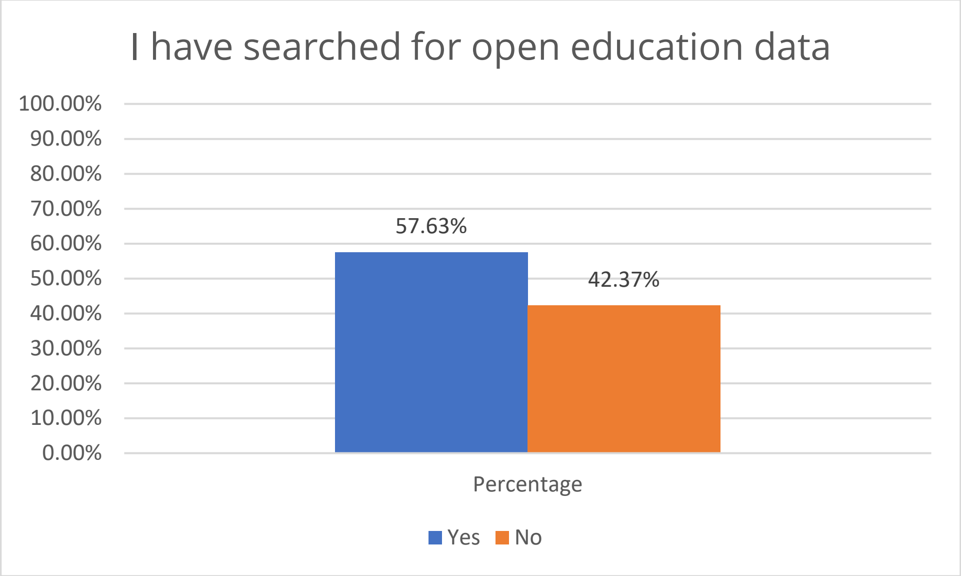


Figure 10 Respondents experience on searching open education data

Figure 10 shows that 57.63% of respondents have experienced on searching for open education data and 42.37% of them never search for open education data.

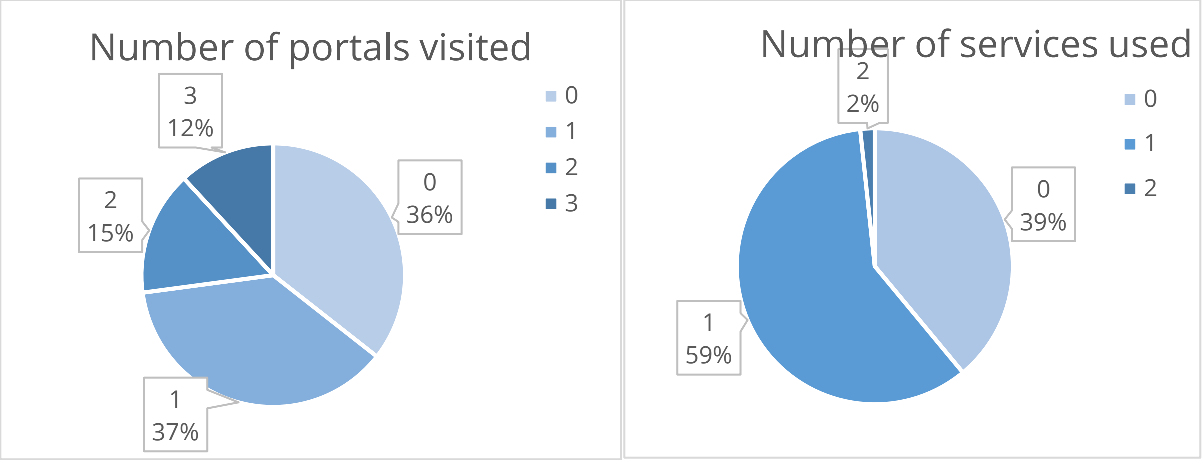


Figure 11 Number of portals visited, and the number of services used

The number of portals visited is counted from a question about education data portal that the respondents have previously visited. The choices are data.overheid.nl portal, OCW portal, DUO portal, Education Inspection agency portal, and VSNU portal. The respondent can also add another open data portal that they have visited before, some of the respondents add CBS and World Bank portal.

Figure 11 shows that 64% of the respondents have visited at least one open education data portal and 36% of them never visited open data portal at all. 27% of the respondents have visited more than one open education data portal, 15% visited two open education data portals, and 12% visited three open education data portals.

The number of services used is counted from a question about open data services that the respondents have previously used. The choices are studiekeuze123.nl and scholenopdekaart.nl. The respondent can also add other open data services that they have used before, some of the respondents add studeersnel.nl.

61% of the respondents have used at least one service created from open education data (i.e., studiekeuze123.nl or scholenopdekaart.nl), and 39% of them never use any services created from open education data.

However, the majority of the respondents (93%) never attend or involve in open education data events, and only 7% of them have attended at least one event. The finding shows that majority of the respondents have previous experience on searching and using open education data portal and services created from open education data, but less experience regarding open education data events.

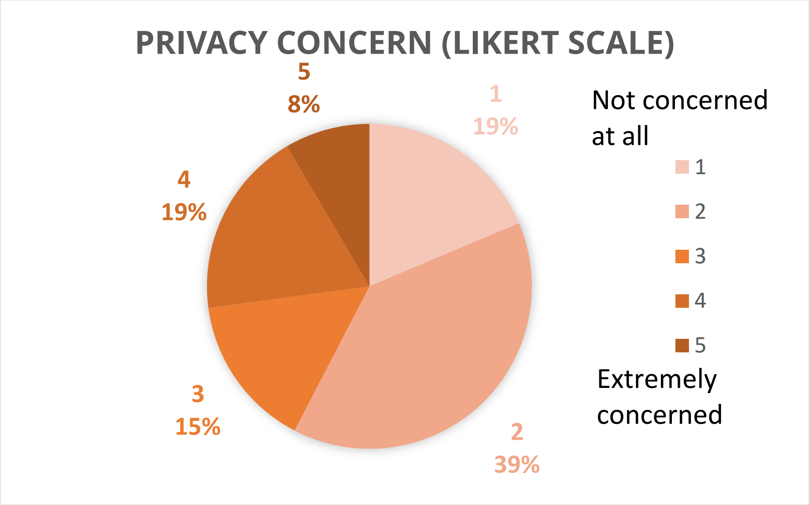


Figure 12 Respondent's privacy concern regarding open education data breach

Furthermore, the respondent's perception towards possible data privacy breaches from open education data is shown in Figure 12. 15% of the respondents are neutral on their reaction, 58% of them are not really concerned, and 27% of the respondents are extremely concerned about data privacy breach. The question includes the description of the extent of open education data breach as follow: "*The open education data is anonymized. The personal data leakage happens when a person can be identified by the combination of multiple anonymous open datasets*".

Respondents most and least important attributes

The respondents are also asked about the most and least important attribute when they choose between two open education data policy alternatives.

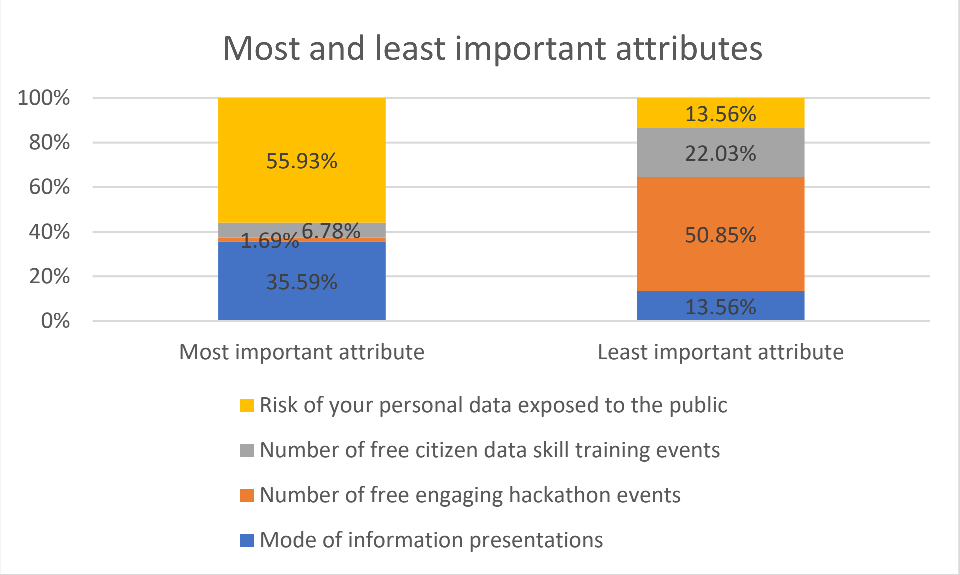


Figure 13 Most and least important attributes

Figure 13 shows that *"risk of your personal data exposed to the public"* and *"mode of information presentation"* are two of the most important attributes for the respondents. 55.93% of the respondents chose *"risk of your personal data exposed to the public"* and 35.59% chose *"mode of information presentation"*. The least important attributes for the respondents are *"number of free engaging hackathon events" and "number of free citizen data skill training events"* with the distribution of 50.85% and 22.03% respectively.

Respondents choice distributions

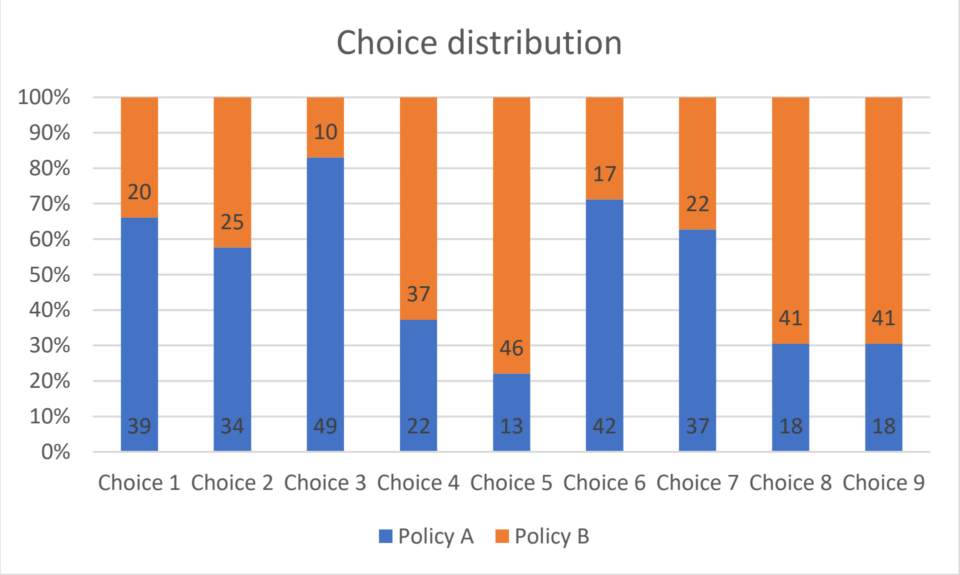


Figure 14 Choice distribution

Figure 14 shows the choice distribution among the respondents; it is not quite balanced for Choice 3, Choice 5, Choice 6, Choice 8, and Choice 9. All these choice situations show that the dominant choices are policies with a lower risk of data leakage. Choice 1, 2, 4, and 7 give a quite balance choice distributions because the risk of data leakage is the same between the two alternatives.

The choice distributions show that respondents highly valued policy with better data protection and hardly willing to trade it with other attributes. In section 5.3, the utility value derived from each attribute by the citizen will be discussed to understand the extent of this non-trading behavior.

Respondents perception to the experiment

The next analysis assesses how the respondents perceived the different surveys. Table 13 shows the answer distributions to the statements about the difficulty, realism, and relevance of the survey. The statements are:

Table 13 Respondents perception on the survey

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Average Score
I was frequently convinced of my choice (1 = strongly disagree, 5 = strongly agree)	6.78%	20.34%	30.51%	38.98%	3.39%	3.12
I think the choice situations are realistic (1 = strongly disagree, 5 = strongly agree)	3.39%	13.56%	57.63%	23.73%	1.69%	3.07
This experiment provides relevant information for the Government to make decisions (1 = strongly disagree, 5 = strongly agree)	8.47%	8.47%	28.81%	47.46%	6.78%	3.36

The first statement asked whether the respondents are convinced of their choice in the survey. 38.98% of the respondents agree, 30.51% of them are neutral, and 20.34% of the respondents disagree with this statement.

The second statement asked whether the choice situations are realistic or not and the majority of the respondents 57.63% are neutral. Therefore, it is understandable that in the first statement 27.12% of the respondents are not convinced of their choice.

In the final statement, the respondents are asked whether the information collected from the survey is relevant for the government to make decisions. 54.24% of the respondents (strongly) agree with the statement, 28.81% of them are neutral and 16.94% (strongly) disagree.

The average scores for the first, second, and third statement are 3.12, 3.07, and 3.36 respectively. Mouter et al., (2017b) in their study regarding citizens trade-off between travel time and safety present similar statements. The respondents in the study show the higher average score for these three statements. The respondents convinced of their choice with a score of 4.5, perceived the choice as realistic with a score of 3.5, and believe the experiment provide relevant information for the government with a score of 3.6

The respondent perception of the experiment shows that the result of this experiment should be taken with careful consideration. It can happen due to the context of the survey (open education data policy) and the attributes that might not be familiar for all the respondents (Mode of information presentation, number of hackathon and training events, data leakage incidents). Compared to the Mouter et al., (2017b) study that asked respondents to choose between road projects with travel time and safety as the trade-off attributes. The respondents may not have the complete information about the extent of open education data policy implementation and face difficulty in measuring the realism of the survey and being confident on their choices.

5.3. Model results

In this section, the survey result is modeled as the MNL (Multinomial Logit) model to estimate the relative values of open education data attributes for the respondents. The MNL model is suitable for the goal of this research which is to estimate the citizen preferences of open data policy attributes.

The model can be used to gain insight into the main effect of each attribute toward the citizen perceived utility. The MNL model on the probability of individual i choosing alternative q is shown in the following equation:

$$Piq = P\ (i\ |Cq) = \frac{e^{V_{iq}}}{\sum_{j \in Cq} e^{V_{iq}}}$$

Where:
Piq is the probability an individual i chooses alternative q
Viq is the utility of individual i to choose alternative q
Cq is the choice set of j alternatives for individual i

Its simple mathematical representation and ease of use also aid the comprehensibility of the model. The MNL model has a disadvantage since it assumes homogenous preferences for a sample which lead to a model with the low goodness of fit and prediction capability. However, even with the low goodness of fit, the model is still useful to estimate the utility values derived from each parameter.

The MNL model parameters are specified in Table 14. The utility parameters for *“number of free engaging hackathon events”* (B_Whackathon), *“number of free citizen data skill training events”* (B_Wtraining), and *“risk of your personal data exposed to the public”* (B_Wprivacy) are estimated linearly. The attribute *“mode of information presentation”* (*B_Wdata*) is dummy coded where the levels represent the complexity of implementation. The dummy coding scheme is sketched in Table 15.

Table 14 MNL Model Parameters Specification

MNL Model Parameter Specification	
Variable	Parameter
B_Wdata_raw	*βdata_raw*
B_Wdata_figures	*βdata_figures*
B_Wdata_services	*βdata_services*
B_Whackathon	*βhackathon*
B_Wprivacy	*βprivacy*

B_Wtraining	*βtraining*

Table 15 Dummy coding for attribute "mode of information presentation"

	β_DATA_RAW	β_DATA_FIGURE	β_DATA_SERVICE
Level 2: Data as services	0	0	1
Level 1: Data as figures	0	1	0
Level 0: Data in original form	1	0	0

Several hypotheses for the signs of the utility parameters are set up. First, the negative estimate sign is expected for the attribute "*risk of your personal data exposed to the public"*. It is expected that the increasing data breach from 1 incident per year until 1 incident per month will result in a decrease in a respondent's utility for an alternative.

Second, positive estimate signs for *"number of free engaging hackathon events"* and *"number of free citizen data skill training events".* Increasing number of participatory & engagement events will increase the utility derived by respondents from an alternative.

Finally, a positive estimate sign with the non-linear utility is expected for *"mode of information presentation"*. The attribute levels are represented in ordinal values. Hence the utility value derived from each attribute level cannot be estimated linearly.

Model estimates

Table 16 summarizes the model estimations for the experiment. Denote that all statistically significant attributes have their a priori expected sign. Furthermore, the estimation parameter and p-value of the statistically significant utility parameters (on a 95% confidence interval) are highlighted in red.

Table 16 Model estimates without checking for linearity

Observations	531			
Individuals	59			
Rho-square	0.121			
Variable	Estimation	Standard Errors	t-test	p-value
βDATA	0.332	0.0935	3.55	0
βHACKATHON	0.0352	0.113	0.31	0.75
βPRIVACY	-0.702	0.0903	-7.78	0
βTRAINING	0.0748	0.0947	0.79	0.43

The statistically significant attributes are *"mode of information presentation"* (B_Wdata) and *"risk of your personal data exposed to the public"* (B_Wprivacy). The "*risk of your personal data exposed to the public"* is statistically significant with an estimation parameter of -0.702; the attribute is estimated linearly. Therefore, this means for "*risk of your personal data exposed to the public"* attribute an increase of incident frequency from 1 incident per year to 1 incident per quarter will reduce the utility of an alternative by 0.702.

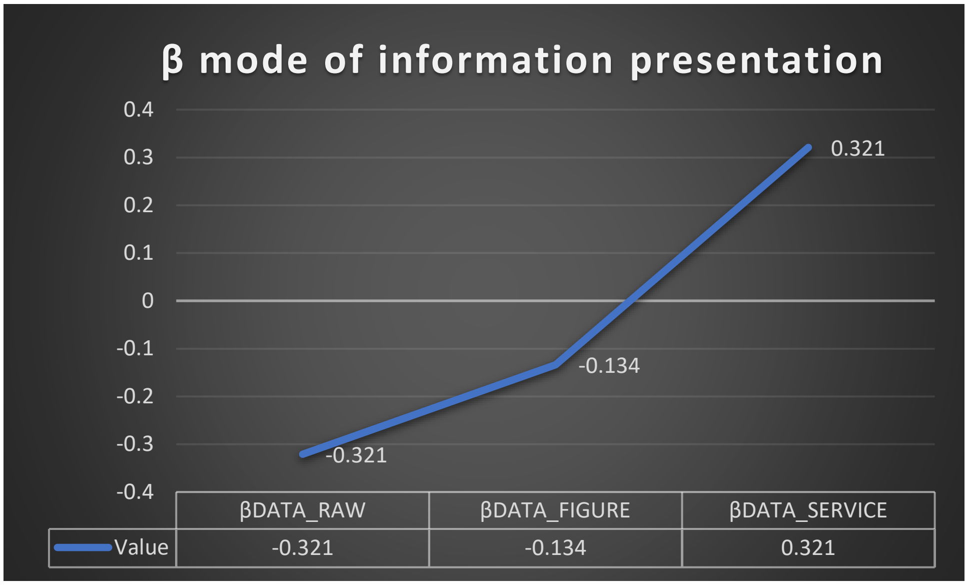


Figure 15 Mode of information presentation utility

Figure 15 shows the utility of *"mode of information presentation"* which are coded into (β_DATA_RAW, β_DATA_FIGURE, and β_DATA_SERVICE). It shows that changing the mode of data presentation from raw data to static/dynamic figures, increase the utility of an alternative by 0.187 (the difference between -0.321 and -0.134). Significant improvement of the utility is shown when the data is presented as a service with 0.455 utility gain from data presented as a figure and 0.642 utility gain from data presented in an original form.

Goodness of fit

The McFadden's Rho-squared statistic is typically measured to evaluate the model fit. The Rho-squared expresses the level of uncertainty the model reduces, compared to a model with all zero estimations. The rho-square of 0.121 signifies that the estimated model can reduce the level of uncertainty by 12.10%, compared to a model with all zeros. Therefore, the MNL model's ability to predict citizen choices between the alternatives is arguable. However, the model is still suitable to identify statistically significant attributes.

5.4. Qualitative results

At the end of the questionnaire, a text box is provided so respondents can give their feedback to improve the questionnaire. The comments are presented in this section.

There is a comment on the realism of "*number of free citizen data skill training events"* attributes. The respondent comments on the alternative implementation of the attributes that are more convenient to reach many audiences (online learning environment).

> *"Would it not be far more convenient for a lot of people to create, for instance, an online learning environment for people to get acquainted with open data?"*

One respondent comments on the realism of "*risk of your personal data exposed to the public"* attribute levels given the effort that the government agencies take to anonymize the data. The attribute levels presented in the survey is perceived as higher than the chance of data leak in reality.

> *"I reckon that much of the open data would be hard to personalize again. I imagine that the chance of leaks is smaller than is proposed here also taking into account the GDPR."*

Other than that, the respondent highlights the importance of marketing efforts as a chance to promote the utilization of open data. The respondent suggests the introduction of the topic from the early education.

> *"Most people likely never heard of training events and many of the hackathons are also likely new to people. I reckon that marketing would be better if people are made enthusiastic at elementary schools and high schools rather than when they are mature already."*

Another respondent suggests to mention the cost of implementation explicitly. In the current survey design, the respondents are asked about their preference by ignoring the fact that every measure taken (more training, hackathon, better security) come with a price.

> *"I think some sort of costs should be included. Of course, no one likes their personal data to be leaked, so people will probably tend to choose for the safest options. The more interesting thing here to know, especially for the government, what is the willingness to pay for certain measures? more trainings, more hackathons, and better security comes with a price, but given that it is not included in the choice set, as a respondent we have to ignore that fact."*

> *"Explain what we have to take into account should we know that they use our money to provide these things. If this is the case, then maybe we should know what the costs are"*

5.5. Conclusion

The descriptive results show that the majority of the respondents is familiar with open education data portals and the services created from open education data. 64% of the respondents have visited at least 1 open education data portal, and 61% of the respondents have used at least 1 service created from open education data. However, only 7% of the respondents have attended open education data events.

The majority of the respondents (73%) is not really concerned about the possibility of data privacy breach from open education data. However, the model result shows that if the data breach incident happens, the respondents are critical to the impact of a data breach on the utility that they gain from the open education data policy. The impact can offset the utility gain from the improvement of other open data policy attributes and dominate their choices.

Answering the sub research question posed at the beginning of this chapter: *What is the valuation of each trade-off attributes for the respondents in their role as a citizen?*

The model result shows two of the most significant attributes are *"mode of information presentation"* and *"risk of your personal data exposed to the public"*. The "*risk of your personal data exposed to the public"* is statistically significant with an estimation parameter of -0.702 which means that an increase of incident frequency from 1 incident per year to 1 incident per quarter will reduce the utility of an alternative by 0.702. The *“mode of information presentation”* is a non-linear attribute significant improvement of the utility is shown when the data is presented as a service with 0.455 utility gain from data presented as a figure and 0.642 utility gain from data presented in an original form.

The attributes *“number of free citizen data skill training events”* and *“number of free engaging hackathon events”* are insignificant in the model. This result is congruent with the most and least important attributes from the descriptive results which show that the respondents are consistent in their choices. However, the descriptive result shows that only 7% of the respondents have attended open education data events. It might have been difficult for respondents to assess their preferences for participation and engagement events (hackathon/data skills training) if they have never attended one.

Chapter 6: Implications and recommendations for policymakers

This chapter is aimed to answer sub research question 5: *Considering the citizen preferences results, what are the recommendations to policymakers creating the Dutch open education data policy?*

In this chapter, the implication of citizen preferences for open education data policy, and the recommendations for policymaker will be discussed. In the first part, the assumptions and limitations of the study are presented for the policymakers consideration in the interpretation of the results. After that, the citizen preferences for open education data policy attributes and the scenario analysis is presented. Finally, the recommendations for policymakers to improve open education data policy are discussed.

6.1. The assumptions and limitations of the study

Before discussing the implications of the results, there are several assumptions and limitations in this study. First, the target respondents for this study is limited to Dutch higher education students and the content of the survey is design to fit their context. Therefore, the result of this study is based on the preferences of Dutch higher education students in their role as a citizen.

Second, the descriptive result shows that only 7% of the respondents have attended open education data events. It can be a reason for hypothetical bias (respondents choose attributes that are familiar to them). Replicating this study with a more balanced sample of respondents (who have experienced all the attributes presented in the questionnaire) will give a better insight on whether the respondents have a true strong preference for *"risk of your personal data exposed to the public"* and *"mode of information presentation"* and not from alternative explanations (e.g., misunderstanding, boredom, strategic behaviour).

Third, this research is the first attempt to empirically measures citizens preferences for open education data policy attributes and by no means set the definitive valuation of trade-off attributes discussed in this study. I believe that the attributes estimation obtained in the study are reasonable and reflect Dutch higher education students' preference for open education data policy in their role as a citizen. However, further replication of the study with more diverse respondents is needed for conclusive valuation of attributes presented in this study. The result of this study should become the basis for further academic discussion and investigation.

6.2. The citizen preferences for open education data policy attributes

The result of citizen stated choice experiment shows citizens' significant preference for *"mode of information presentation"* and *"risk of your personal data exposed to the public"*.

Observations	531			
Individuals	59			
Rho-square	0.121			
Variable	Estimation	Standard Errors	t-test	p-value
βDATA	0.332	0.0935	3.55	0
βHACKATHON	0.0352	0.113	0.31	0.75
βPRIVACY	-0.702	0.0903	-7.78	0
βTRAINING	0.0748	0.0947	0.79	0.43

The *"risk of your personal data exposed to the public"* is a linear attribute and has the highest utility estimation of -0.702. It means that each movement of attribute levels will reduce the utility of an alternative by 0.702. There are three attribute levels: 1 incident per year, 1 incident per quarter, and 1 incident per month. It means a policy with 1 incident per year is valued 0.702 more than a policy with 1 incident per quarter and valued 1.404 more than a policy with 1 incident per month attribute levels.

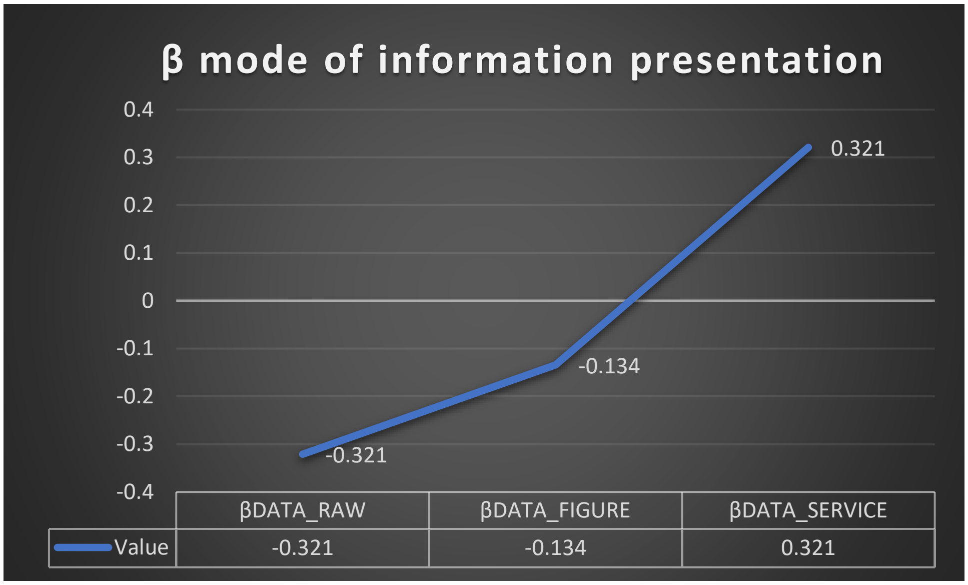


Figure 16 Recap mode of information presentation utility

The *"mode of information presentation"* is a non-linear attribute with a slight difference of 0.187 utility estimation between the information presented in the original form and the static/dynamic figure. There is a significant utility gain if the information is presented as a service compared to other attribute levels, 0.455 utility gain over information presented in the figures and 0.642 utility gain over information presented in the original form. However, the gain is not enough to offset the utility reduction from *"risk of your personal data exposed to the public"* which can explain the dominant alternative in several choice situations.

The attributes *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"* are insignificant in the model. This result is congruent with the most and least important attributes from the descriptive results which show that the respondents are consistent in their choices. However, the descriptive result shows that only 7% of the respondents have attended open education data events. It might have been difficult for respondents to assess their preferences for participation and engagement events (hackathon/data skills training) if they have never attended one.

6.3. Scenario analysis

In the previous section, two of the most significant attributes are identified: *"risk of your personal data exposed to the public"* and *"mode of information presentation".* In this section, two scenarios are developed to illustrate the effect of these two attributes.

Table 17 Reference scenario

Attributes	Attribute levels
Mode of information presentation	in original form (as similar as possible to the source)
Number of free engaging hackathon events	1 every 2-years
Number of free citizen data skill training events	1 per year
risk of your personal education data exposed to the public	1 incident per year

Scenario 1 (service creation): Government developed a policy to focus on services creation and commissioned several external parties. There is an increased chance of data leak because the information is passed to an external party.

Scenario 2 (data protection): Government create a policy that strongly protects the private data. Several layers of approval are needed before the data can be published which leads to the limited supply of education data and the data is published in the original form.

Both scenarios have the same attribute levels for "number of free engaging hackathon events" and "number of free citizen data skill training events" attributes.

Table 18 Scenario analysis

Attributes	Reference scenario	Scenario 1 (service creation)	Scenario 2 (data protection)
Mode of information presentation	in original form (as similar as possible to the source) Utility = - 0.321	as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl) Utility = 0.321	in original form (as similar as possible to the source) Utility = - 0.321
Number of free engaging hackathon events	1 every 2-years Utility = 0.061	1 every 2-years Utility = 0.061	1 every 2-years Utility = 0.061
Number of free citizen data skill training events	1 per year Utility = 0.0954	1 per year Utility = 0.0954	1 per year Utility = 0.0954
risk of your personal education data exposed to the public	1 incident per year Utility = 0.702	1 incident every 3-months Utility = 0	1 incident per year Utility = 0.702
Scenario Total Utility	0.5385	0.4785	0.5385

In the scenario 1 the utility gain from presenting the information as a service compared to the original form is 0.642. However, this gain is offset by the utility loss of -0.702 from the increased chance of data leak from 1 incident per year to 1 incident every 3-months. It makes scenario 1 has a lower total utility than scenario 2 even though the mode of information presentation of scenario 1 is significantly better than scenario 2.

This scenario analysis shows the citizens give a significant negative response to any policy which compromises their personal data protection. The respondents do not want to trade their personal data protection for any improvement in the other attributes. It gives the government agency limited choices to improve the open data policy because the risk for opening data and compromise the data privacy is higher for them than the benefits that the other attribute can deliver.

However, given the description of open education data breach as follow: "*The open education data is anonymized. The personal data leakage happens when a person can be identified by the combination of multiple anonymous open datasets*". 58% of the respondents show no concern about the possibility of data privacy breach. It seems in reality respondents have great trust in the government to protect their privacy, and the wording of choice situation exaggerate the possibility of a data leak. The policymaker should consider this fact in the interpretation of this study, and further investigation is needed to determine the utility of data protection attribute conclusively.

6.4. Recommendations for policymakers

Even though in current conditions government agencies are pressured with limited resources (personnel and monetary) and the requirement to comply with General Data Protection Regulation.

There are several recommendations for the government to improve the open education data policy:

1. Collaborate with infomediaries to provide services for citizens
   The citizens derived significant utility from the information that is provided as a service compared to the other forms (original data and figures). However, there is a lack of services created from open education data.

   Welle Donker & van Loenen (2017) in their investigation on Dutch open education data shows that there is a strong desire from infomediary users to build a partnership with government given the right stimulation (monetary compensation).

   It is important for the government to build a partnership with infomediary users who use the raw open education data to create functional services for other citizens. The government can collect the requirements for new services during its annual event "OCW Kennisfestival" and consult with the respective stakeholders (education council, students, parents) afterward for the detail specifications. After that, the government can commission the creation of the service to the infomediary users. In this process, the government can use monetary incentives to motivate the infomediary users for creating the service.

   Commissioning the service creations to infomediary users also enable the services to use the education data that are not publicly available. The government can provide those data directly to them and control the handling of the data. For example, studiekeuze123.nl have non-public data from the National Student Survey that it uses for measuring student satisfaction in the study program. The non-public data may have useful information that cannot be disclosed according to the privacy assessment model in the common open education data. Other than that, citizens are more likely to trust and use services that are officially commissioned by the government.

The Ministry of Education, Culture and Science have enough experience in this schema of partnership with infomediary users as can be shown from the "Windows for Accountability" project which leads to the creation of scholenopdekaart.nl and studiekeuze123.nl.

The schema also produces the highest possible utility combination because the government can provide the information in the form of services while protecting the citizens personal data internally. However, given the limited budget of the government, there will be a trade-off between this recommendation and immediate needs to comply with the GDPR requirements.

2. Engage the citizens in an effective and cost-efficient manner
The two least significant attributes are *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"*. Between these attributes, the citizens prefer *"number of free citizen data skill training events"* over the *"number of free engaging hackathon events"*. It can be interpreted that citizens prefer the improvement of the data literacy of the general population rather than the one-time event such as a hackathon.

   It is expected because the majority of the citizens are not interested in participating in the hackathon events, but they may perceive data skill training events as more beneficial for the general population.

   The data skills training can be implemented in different forms:
   - One of the respondents recommends creating an online course that can be freely accessed by the citizens.

     > *"Would it not be far more convenient for a lot of people to create, for instance, an online learning environment for people to get acquainted with open data?"*

   - Another respondent recommends engaging the students from the early level of education. The government can embed the data literacy skills in the education curriculum as well.

     > *"Most people likely never heard of training events and many of the hackathons are also likely new to people. I reckon that marketing would be better if people are made enthusiastic at elementary schools and high schools rather than when they are mature already.*

6.5. Conclusion

Answering the sub research question posed at the beginning of this chapter: *Considering the citizen preferences results, what are the recommendations to policymakers creating the Dutch open education data policy?*

Citizens highly preferred *"mode of information presentation"* and *"risk of your personal data exposed to the public"* as the open education data policy attributes. There is a significant difference of utility between different mode of information presentation, providing the information as a service improve the utility gain by 0.642 and 0.455 compared to the information in the original form and in the figures respectively.

However, the scenario analysis shows that the utility gained from *"mode of information presentation"* attribute cannot offset the reduction of utility caused by *"risk of your personal data exposed to the public"* attribute. Changing 1 incident of data leak per year into 1 incident of data leak per quarter results in the reduction of utility by 0.702 which offset the highest utility gain of the *"mode of information presentation"* attribute. It explains the existence of dominant choice in the descriptive result of choice distribution.

Given the description of open education data breach as follow: "*The open education data is anonymized. The personal data leakage happens when a person can be identified by the combination of multiple anonymous open datasets*". From the descriptive result, 58% of the respondents show no concern about the possibility of data privacy breach. It seems in reality respondents have great trust in the government to protect their privacy, and the wording of choice situation exaggerate the possibility of a data breach. The policymaker should consider this fact in the interpretation of this study, and further investigation is needed to determine the utility of data protection attribute conclusively.

The attributes *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"* are insignificant in the model. However, the descriptive result shows that only 7% of the respondents have attended open education data events. It might have been difficult for respondents to assess their preferences for participation and engagement events (hackathon/data skills training) if they have never attended one.

Two recommendations are formulated for the policymakers:

1. Collaborate with infomediary to provide services for citizens
2. Engage the citizens in a cost-efficient and subtle manner

Chapter 7: Discussion and Conclusion

This research assessed which factors affect the preferences of citizens for the Dutch open education data policy. These preferences were elicited through the design of a discrete choice experiment. Each of the section in the following part will address a specific research sub-questions formulated in Chapter Research Design.

7.1. Dutch open education data policy context

The following sub-question is addressed in Chapter 1: Introduction:

> *"What is the policy context (policy objectives, organization, existing implementation) of Dutch open education data policy?"*

OCW policy objective is 'education quality openness,' with the future vision as follow (Rijksoverheid, 2018b):

- Parents and pupils know where they can find important information about schools.
- Parents and pupils use this information to compare schools and choose a suitable school.
- Parents, pupils, and the education council use the information to discuss the quality of education with the school.
- All schools use the available data in the best possible way to improve education.
- All government data is public and is used to develop useful applications for parents, students, teachers, and school leaders.

Three government agencies oversee the implementation of open education data (DUO, OCW, and Education Inspection Agency). DUO as the executive agency is in charge of processing and publishing open education data held by OCW.

Based on the policy context exploration, there is the significant implementation of data-related attributes and participation & engagement attributes within Dutch open education data policy. The OCW provides information in diverse forms such as raw data in the respective open data portals (DUO, OCW, Education Inspection Agency), static and interactive figures (OCW and VSNU portals) and creating services from open education data (scholenopdekaart.nl and studiekeuze123.nl).

Other than that, several participation & engagement events are organized such as data exploration event "Education Data under scrutiny" and hackathon "Hack de Valse Start". There is no specific portal-related attributes implementation in the OCW open education data policy; all the data are simply hosted in each agency open data portal without any additional features for the users to interact with the portal (data visualization/data analysis tools).

On 25th May of 2018, the General Data Protection Regulation (GDPR) is formally applied in the Netherlands. The introduction of GDPR reinforces the existing barrier faced by government agencies in opening their data (risk-averse culture and limited resource to handle the data publishing process). The risk of opening data is increased because there is a hefty fine in case of data breaches (as high as €20 million or €10 million according to the bill).

In order to comply with the data protection specification of the GDPR, sizeable resources are required (both human resources and monetary) which will put pressure on their budget for other functions. OCW hires two Data Protection Officers, one at DUO and one at the board department. A specific FG at DUO was chosen because of the large amount of personal data at DUO and the need to exercise adequate supervision at a short distance (OCW, 2017). The Data Protection Officer is in charge of Data Protection Impact Assessments (DPIA), mapping the privacy risks of a data processing system in advance and take measures to reduce the risks.

7.2. Identification of potential trade-off attributes from the literature

The following sub-question is addressed in the literature review from Chapter 3: Open Data Policy Preference Study: State of The Art:

> *"What are the possible trade-off attributes for the open data policy in the existing literature?"*

The literature review identifies a tension between 'stewardship' and 'usefulness' principles in the open data policy. This tension is mapped as data-related and participation/collaboration-related capabilities/processes in Open Government Maturity Model (OGMM) by Lee & Kwak (2012). Furthermore, the existing open data policy assessment study approaches the open data policy from diverse perspectives which are open data portal perspective, the socio-technical perspective, and citizen perspective.

Three common categories of open data policy attributes are identified from the literature review: data-related attributes, portal-related attributes, and participation & engagement related attributes.

- **Data-related attributes** consist of data availability, data quality, data discoverability, and data protection.
- **Portal-related attributes** are communication and interaction features, open data portal ease of use.
- **Participation & engagement related attributes** are public awareness, public participation, motivation, development of required skills and expertise, compatibility of the data provided with the needs, and data reusability.

The attributes are selected based on its possibility to be directly experienced by the citizens. If the citizens have experience related to the attributes it will help them to understand the survey and give a valid response. Therefore, attributes that are hardly perceived by the citizens and related to the internal arrangements of data providers are excluded. For example, vision and leadership, organization restructuring, interagency communication, legislation.

7.3. Design of Citizen Stated Choice Experiment

The following sub-question is addressed in Chapter 4: Citizen stated choice experiment design:

> *"How do the identified trade-off attributes and policy context translate into the citizen stated choice experiment design?"*

The citizen stated choice experiment is designed based on three categories of attributes identified in Chapter 3: data-related attributes, portal-related attribute, and participation & engagement attributes. Based on the policy context identified in Chapter 1 the portal-related attribute is omitted because there is the limited implementation of open education data portal. The open education data is simply hosted in the respective government agency portal (DUO, Education Inspection Agency, OCW) without any features for user interaction (visualization, data analysis).

However, there is an increasing importance for one of the data-related attributes which are the data protection. It emerges as a significant attribute due to the passing of the General Data Protection Regulation (GDPR) in 25th May of 2018. Government agencies face increasing barriers (risk and limited resource) in opening data.

The final selection of attributes are data-related attributes, data protection attribute, and participation & engagement attributes. Each of the attributes is further specified into measurable options, *"mode of information presentation"* for the data-related attributes, *"risk of your personal education data exposed to the public"* for the data protection attribute, and *"Number of free engaging hackathon events"* and *"Number of free citizen data skill training events"* for the participation & engagement attributes.

These attributes are then used to generate a fractional factorial orthogonal design with nine choice situations. Basic plan 2 design is chosen, with three attributes in three levels and a total of 9 choice sets. The experiment is generated using Ngene software with a sequential construction of the alternatives. The recap of attribute levels and values used to generate the choice sets is shown in Table 19 .

Table 19 Recap attribute levels and values

Category	Attributes	Value
Data-related attribute	Mode of information presentation	• in original form (as similar as possible to the source) • as static or interactive figures • as a service (e.g., an application such as studiekeuze123.nl or scholenopdekaart.nl)
Participation & engagement related attribute	Number of free engaging hackathon events	• 1 every 2-years • 1 per year • 2 per year
	Number of free citizen data skill training events	• 1 per year • 2 per year • 3 per year
Data protection attribute	risk of your personal education data exposed to the public	• 1 incident per year • 1 incident every 3-months • 1 incident per month

After that, the pilot survey consists of three parts are constructed: 1) Leading questions about open education data policy, 2) Choice situations, and 3) Perception and demographic questions. The pilot survey is tested among ten respondents, and the feedbacks are incorporated in the final survey.

The final survey is improved based on the following feedbacks: 1) include the description for each attribute, 2) reduce the wordiness of choice situations, and 3) clearly define the extent of data leakage.

7.4. The result of Citizen Stated Choice Experiment

The following sub-question is addressed in Chapter 5: Citizens preferences for a Dutch open education data policy:

> *"What is the valuation of each trade-off attributes for the respondents in their role as a citizen?"*

Hypothesis 1: negative estimate sign for the attribute "*risk of your personal data exposed to the public*"

Hypothesis 2: positive estimate signs for *"number of free engaging hackathon events"* and *"number of free citizen data skill training events"*

Hypothesis 3: positive estimate sign with non-linear utility for *"mode of information presentation"*

For attribute "*risk of your personal data exposed to the public*", it is expected that the increasing data breach from 1 incident per year until 1 incident per month will result in a decrease in a respondent's utility for an alternative.

Increasing number of participatory & engagement events will increase the utility derived by respondents from an alternative. For *"mode of information presentation",* it is expected that the change from basic mode of information (data provided in original form) to the next attribute level (data provided as services) will increase respondent's utility for an alternative. The attribute levels are represented in ordinal values. Hence the utility value derived from each attribute level cannot be estimated linearly.

Table 20 Recap model estimates without checking for linearity

Observations	531			
Individuals	59			
Rho-square	0.121			
Variable	Estimation	Standard Errors	t-test	p-value
βDATA	0.332	0.0935	3.55	0
βHACKATHON	0.0352	0.113	0.31	0.75
βPRIVACY	-0.702	0.0903	-7.78	0
βTRAINING	0.0748	0.0947	0.79	0.43

The model result in Table 20 shows that all of the trade-off attributes have the expected signs of the hypotheses. Positive estimate signs for *"mode of information presentation"*, *"number of free engaging hackathon events",* and *"number of free citizen data skill training events"* attributes*.* Negative signs for "*risk of your personal data exposed to the public"* attribute.

Two of the most significant attributes are *"mode of information presentation"* and *"risk of your personal data exposed to the public"*. The "*risk of your personal data exposed to the public"* is statistically significant with an estimation parameter of -0.702 which means that an increase of incident frequency from 1 incident per year to 1 incident per quarter will reduce the utility of an alternative by 0.702.

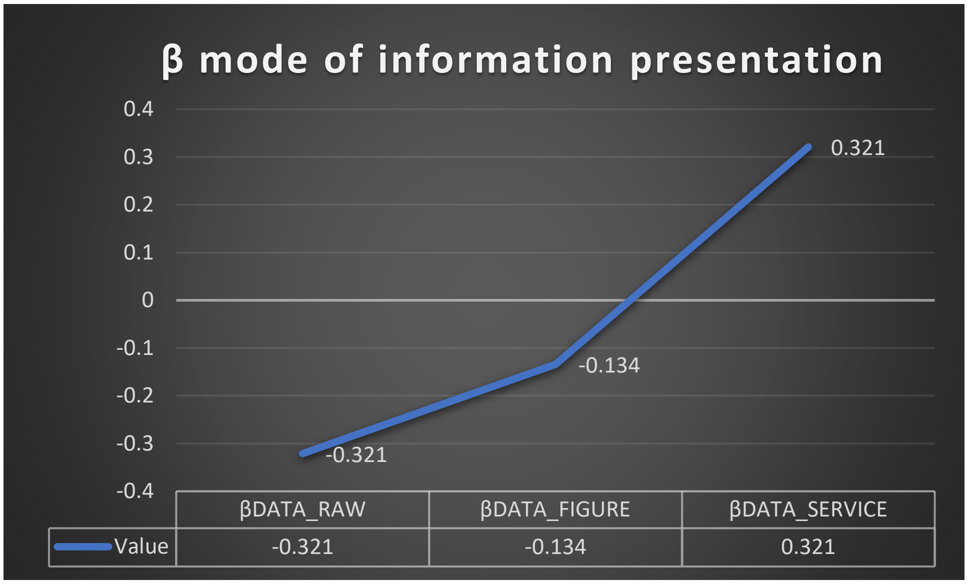


Figure 17 Recap mode of information presentation utility

The *"mode of information presentation"* is a non-linear attribute as shown in Figure 17, significant improvement of the utility is identified when the data is presented as a service with 0.455 utility gain compared to data presented as a figure, and 0.642 utility gain compared to data presented in an original form.

The attributes *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"* are insignificant in the model. This result is congruent with the most and least important attributes from the descriptive results which show that the respondents are consistent in their choices.

The descriptive result shows that majority of the respondents are familiar with open education data portal and the services created from open education data. 64% of the respondents have visited at least one open education data portal, and 61% of the respondents have used at least one service created from open education data. However, only 7% of the respondents have attended open education data events.

Assumptions and limitations for the model interpretation

There are several assumptions and limitations for the interpretation of the result.

First, the target respondents for this study is limited to Dutch higher education students and the content of the survey is designed to fit their context. Therefore, the result of this study is based on the preferences of Dutch higher education students in their role as a citizen.

Second, the descriptive result shows that only 7% of the respondents have attended open education data events. It can be a reason for hypothetical bias (respondents choose attributes that are familiar to them). Replicating this study with a more balanced sample of respondents (who have experienced all the attributes presented in the questionnaire) will give a better insight on whether the respondents have a true strong preference for *"risk of your personal data exposed to the public"* and *"mode of information presentation"* and not from alternative explanations (e.g., misunderstanding, boredom, strategic behaviour).

Third, this research is the first attempt to empirically measures citizens preferences for open education data policy attributes and by no means set the definitive valuation of trade-off attributes discussed in this study. I believe that the attributes estimation obtained in the study are reasonable and reflect Dutch higher education students' preference for open education data policy in their role as a citizen. However, further replication of the study with more diverse respondents is needed for conclusive valuation of attributes presented in this study. The result of this study should become the basis for further academic discussion and investigation.

7.5. Implication and recommendation for the policymaker

The following sub-question is addressed in Chapter 6: Implications and recommendations for policymakers:

> *"Considering the citizen preferences results, what are the recommendations to policymakers creating the Dutch open education data policy?"*

Citizens highly preferred *"mode of information presentation"* and *"risk of your personal data exposed to the public"* as the open education data policy attributes. There is a significant difference of utility between different mode of information presentation, providing the information as a service improve the utility gain by 0.642 and 0.455 compared to the information in the original form and in the figures respectively.

The scenario analysis shows that the utility gained from *"mode of information presentation"* attribute cannot offset the reduction of utility caused by *"risk of your personal data exposed to the public"* attribute. Changing 1 incident of data leak per year into 1 incident of data leak per quarter results in the reduction of utility by 0.702 which offset the highest utility gain of the *"mode of information presentation"* attribute. It explains the existence of dominant choice in the descriptive result of choice distribution.

However, given the description of open education data breach as follow: "*The open education data is anonymized. The personal data leakage happens when a person can be identified by the combination of multiple anonymous open datasets*". From the descriptive result, 58% of the respondents show no concern about the possibility of data privacy breach. It seems in reality respondents have great trust in the government to protect their privacy, and the wording of choice situation exaggerate the possibility of a data breach. The policymaker should consider this fact in the interpretation of this study, and further investigation is needed to determine the utility of data protection attribute conclusively.

The attributes *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"* are insignificant in the model. However, the descriptive result shows that only 7% of the respondents have attended open education data events. It might have been difficult for respondents to assess their preferences for participation and engagement events (hackathon/data skills training) if they have never attended one.

Two recommendations are formulated for the policymakers:

1. Collaborate with infomediary to provide services for citizens

   The citizens derived significant utility from the information that is provided as a service compared to the other forms (original data and figures). However, the citizens lack of motivation to contribute for service creation will lead to bottleneck on the creation of new services based on open education data.

   It is important for the government to build a partnership with infomediary users which use the raw open education data to create functional services for other citizens. The government can collect the requirements for new services during its annual event "OCW Kennisfestival" and consult with the respective stakeholders (education council, students, parents) afterward for the detail specifications. After that, the government can commission the creation of the service to the infomediary users. In this process, the government can use monetary incentives to motivate the infomediary users for creating the service.

   Commissioning the service creations to infomediary users also enable the services to use the education data that are not publicly available. The government can provide those data directly to them and control the handling of the data. For example, studiekeuze123.nl have non-public data from the National Student Survey that it uses for measuring student satisfaction in the study program. The non-public data may have useful information that cannot be disclosed according to the privacy assessment model in the common open education data. Other than that, citizens are more likely to trust and use services that are officially commissioned by the government.

   The Ministry of Education, Culture and Science have enough experience in this schema of partnership with infomediary users as can be shown from the "Windows for Accountability" project which leads to the creation of scholenopdekaart.nl and studiekeuze123.nl.

   The schema also produces the highest possible utility combination because the government can provide the information in the form of services while protecting the citizens personal data internally. However, given the limited budget of the government, there will be a trade-off between this recommendation and immediate needs to comply with the GDPR requirements.

2. Engage the citizens in a cost-efficient manner

   The two least significant attributes are *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"*. Between these attributes, the citizens prefer *"number of free citizen data skill training events"* over the *"number of free engaging hackathon events"*. It can be interpreted that citizens prefer the improvement of the data literacy of the general population rather than the one-time event such as a hackathon.

   It is expected because the majority of the citizens are not interested in participating in the hackathon events, but they may perceive data skill training events as more beneficial for the general population.

   The data skills training can be implemented in different forms:

- One of the respondents recommends creating an online course that can be freely accessed by the citizens.

> *"Would it not be far more convenient for a lot of people to create, for instance, an online learning environment for people to get acquainted with open data?"*

- Another respondent recommends engaging the students from the early level of education. Government can embed the data literacy skills in the education curriculum as well.

> *"Most people likely never heard of training events and many of the hackathons are also likely new to people. I reckon that marketing would be better if people are made enthusiastic at elementary schools and high schools rather than when they are mature already."*

7.6. Citizen preferences for an open education data policy in the Netherlands

Finally, the main question is addressed:

> *"What are the preferences of citizens for a Dutch open education data policy?*

Based on the citizen stated choice experiment, the Dutch higher education students in their role as a citizen significantly valuate data protection attribute (*"risk of your personal data exposed to the public"*) and data-related attributes ("mode of information presentation").

Between three type of "mode of information presentation", citizens derive significant value if the data is presented as a service compared to data presented as a figure, and data presented in an original form. However, the value gained from the improvement in "mode of information presentation" is not enough to offset the loss of value in case of a data breach.

Therefore, the government agency has limited choices to improve the open data policy because the risk for opening data and compromise the data privacy is higher for them than the benefits that the other attribute can deliver.

However, the possibility of 'hypothetical bias' should be considered in the interpretation of the result. In the survey, open education data breach is described as follow: "The open education data is anonymized. The personal data leakage happens when a person can be identified by the combination of multiple anonymous open datasets".

58% of the respondents show no concern about the possibility of data privacy breach. It seems, in reality, respondents have less concern about the possibility of data breach and the wording of choice situation exaggerate the chance. The policymaker should consider this fact in the interpretation of this study, and further investigation is needed to determine the utility of data protection attribute conclusively.

Other than that, two attributes are considered insignificant by the citizens *"number of free citizen data skill training events"* and *"number of free engaging hackathon events"*. However, the descriptive result shows that only 7% of the respondents have attended open education data events. It might have been difficult for respondents to assess their preferences for participation and engagement events (hackathon/data skills training) if they have never attended one.

Given the citizens reluctance to compromise the data protection attribute, government agencies have limited option for the implementation. Two recommendations are formulated to improve the existing open education data policy: 1) Collaborate with infomediary to provide services for citizens, and 2) Engage the citizens in a cost-efficient manner.

7.7. Limitation of the study

There are several limitations to the study:

Hypothetical situations instead of real situations

In the discrete choice experiment, the choice situations represent hypothetical situations rather than real situations. Therefore, it remains the question if respondents would make the same choices in a real-life situation.

Characteristics of respondents

The final survey is distributed among students who are currently attending a Dutch higher education institution or recently graduated. The higher education students are targeted due to several reasons: 1) have relevant use case for the open education data which make them more likely to know about open data, 2) have relevant skills to use open education data, and 3) more likely to understand the term used in the survey with a proper explanation. The survey will gather different results if it is distributed in the general population, with more respondents who are not familiar with open data policy. In order to mitigate the homogenous characteristic of the respondents, the survey is distributed to the students with diverse study programs.

A limited number of respondents

The citizen stated choice experiment is distributed to 59 respondents. Each of the respondents completes nine choice situations which result in 531 choice observations. These observations become the basis for Multinomial Logit (MNL) model created in this study.

Limited selection of attributes

The attributes are selected based on three criteria: 1) Expected influence on an individual, 2) Societal relevance of the factor, and 3) Measurability in the discrete choice experiment. The attributes selected for the experiment are limited and may not reflect the whole possibility of attributes for citizens. For example, one of the respondents comments about using data skill training events as one of the attributes while there is another cheaper option such as creating an online learning environment that can be freely accessed by the citizens.

Using secondary source for the policy context exploration

The policy context exploration is conducted through desk research on the published policy documents of the government agencies responsible for open education data policy such as Ministry of Education, Culture and Science (OCW), Education Executive Agency (DUO), and Education Inspection Agency (Inspectie van het Onderwijs). However, there is no primary source in the form of direct communication with those respective agencies because the agencies do not accept the request for an interview for a student project.

Exclusion of cost of implementation

The cost variable is not included in the experiment due to the lack of information regarding the cost of implementation for each attribute from the secondary source. The information from the secondary source is highly aggregated and only shows the budget for the whole government agency. One of the respondents comments about the lack of a cost attribute which will become one of the most important attributes for the respondents to compare between different alternatives.

7.8. Recommendations for future study

Extend the research for different context of open data policy

In this research, the experiment is limited to open education data and higher education students as the target respondents. Future research can explore different policy context (e.g., open data policy for geospatial data, science data) or different respondents for open education data. For example, open education data policy for primary and secondary schools which targets the parents and pupils as the users.

Expand research with unobserved alternatives and attributes

In this research four attributes are used to generate the choice situations. However, in reality, many attributes can be included or combined to make different alternatives. The portal-related attribute is omitted from this study because of the limited implementation of open data portal in the Dutch open education data. However, if the future research explores the portal-related attributes of city open data portal, the attributes selected will be different from the attributes in this study. The attributes will focus on the functionality and features of the open data portal such as the visualization capability, collaboration and communication features, the format of the data provided, compared to the socio-technical perspective of this study.

Validate the result using different models (Mixed Logit Model and Latent Class Analysis)

In this study, Multinomial Logit (MNL) model is chosen for the model estimation. Homogeneity of preference is assumed for the MNL model, and the panel nature of the data are not reflected in the result. Replicating the study with alternative models (Mixed Logit and Latent Class Analysis) could address this limitation.

Mixed Logit Model able to capture the model heterogeneity and accounts for the panel nature of the data; Mixed Logit Model explicitly assumes that there is a distribution of preference weights across the sample reflecting differences in preferences among respondents, and it models the parameters of that distribution for each attribute level (Hauber et al., 2016). Mixed logit model requires assumptions about the distribution of parameters across respondents and larger sample sizes than MNL. Latent Class Analysis (LCA) also able to model the heterogeneity using latent classes which result in parsimonious estimator with a unique solution; it requires smaller samples than Mixed Logit Model (Hauber et al., 2016). However, it requires the assumption to determine an appropriate number of classes to be estimated, and the required sample size varies with the number of classes in the model.

Using primary source information

In the limitation of the study, the exclusive use of the secondary source in the study is discussed. A future study could contact the responsible government agencies and gain access to the primary source information. It is important to improve the realism of the survey and collect detail information that is not publicly available from the published policy documents.

Include trade-off attributes cost of implementation in the survey

If the future research able to secure information from a primary source (interview with government agencies that implement open education data), it is important to include the cost of implementation in the survey. Each attribute implementation certainly comes with a price. However, in this study, the respondents are asked to do a trade-off between attributes without considering the cost of implementation. It will be interesting to investigate whether respondents valuate the trade-off attributes differently if the cost of the implementation is revealed.

It is also interesting to Include different functions of government agencies that require the limited budget in the experiment. In this research, DUO does not have enough budget to implement the changes needed to comply with GDPR requirements unless it compromises the budget for the other functionalities. Do the respondents in their role as a citizen willing to trade-off those functionalities (study loans, reimbursement of school costs, funding educational institutions) with the improvement in open education data attributes?

7.9. Reflection on societal/managerial relevance

The study is conducted to address the problem of “lack of insight into the citizen preferences of open data policy attributes”. This lack of insight has influenced policymakers on how they develop and evaluate open education data policy. In the current situation, government agencies tend to replicate “best practice” policy from other agency without considering their policy objectives and context. This tendency to mimic other agency and lack of insight on the citizens preferences lead them to evaluate and develop their open data policy only from the data provider perspective. Policymakers tend to use the easily measured attributes (quantity of the data published) or using the established benchmark (e.g., open government data readiness).

This research provides an alternative method for governments to evaluate and develop their open data policy alongside the commonly used government/data provider perspective. It enables policymakers to empirically valuate citizen preferences for specific open data attributes based on their choices of several ‘hypothetical’ open data policy. The valuation of attributes and citizen preference is essential because the developed open data policy and subsequent evaluation should be guided from the citizens perspective. Citizens are the end users of open data, and the policy should benefit them because it is the primary goal of opening data.

This study specifically measures the preferences of Dutch higher education students in their role as a citizen for a Dutch open education data policy. Four attributes are selected to develop ‘hypothetical’ open data policy and measure the citizen preferences. These attributes are “risk of your personal data exposed to the public”, “mode of information presentation”, “number of free citizen data skill training events”, and “number of free engaging hackathon events”.

The result shows that Dutch higher education students in their role as a citizen choose “risk of your personal data exposed to the public” and “mode of information presentation” attributes as the most important attributes for open education data policy. They significantly prefer a policy with lower “risk of your personal data exposed to the public” attribute. This preference even outweighs the benefit that the citizens derived from the improvement of “mode of information presentation” attribute (from data in the original form to present data as a service). This strong preference for data protection limit the options for government agencies in developing their open education data policy.

Government agencies have limited resource (personnel and monetary) for their operation. The valuation enables policymakers to understand how citizens valuate specific attributes in comparison to the others and what is the trade-off for the policymaker if they choose one attribute over the other.

The attributes “number of free citizen data skill training events”, and “number of free engaging hackathon events” are not significant for the respondents. However, only 7% of the respondents have attended open education data events. It can be a reason for hypothetical bias (respondents choose attributes that are familiar for them) which make the estimates value of “risk of your personal data exposed to the public” and “mode of information presentation” higher than its true value in reality.

Replicating this study with a more balanced sample of respondents (who have experienced all the attributes presented in the questionnaire) will give a better insight on whether the respondents have a true strong preference for *“risk of your personal data exposed to the public”* and *“mode of information presentation”* and not from alternative explanations (e.g., misunderstanding, boredom, strategic behaviour).

From the study, we can interpret that citizens are not familiar with policymakers efforts for participation and engagement activities. However, citizens participation and engagement are important for the creation of desired public values(transparency, accountability, and economic growth) promised by the open data.

Two recommendations are formulated based on the result. First, collaborate with infomediary to provide services for citizens. The study found two types of users (infomediary who create services from open data for end users, and end users who consume information from services created by infomediary). Policymakers can stimulate the infomediary involvement to create new services based on open data with a suitable incentive (monetary incentive, supporting infomediary community).

Second, engage the citizens in a cost-efficient manner. Several respondents comment about the possibility of a more cost-efficient alternative to engage citizens compared to organizing annual hackathon or data exploration events. For example, creating an online data learning environment or embedding the data literacy skills in the curriculum for the early studies (primary/secondary school).

7.10. Academic reflection

Design of citizen stated choice experiment (CSCE) for open education data policy

Designing CSCE in open data policy context is challenging because the respondents may not have previous experience and knowledge about the topic. Other than that, compared to similar Discrete

Choice Experiment (DCE) study in transport, health, and environment domain, the choice situation for open education data policy are more abstract and less intuitive for the respondents.

Therefore, we need to make a trade-off between providing more explanation (risk of anchoring effect) and less explanation (risk of validity). In the pilot test, the variant of the survey with less explanation is tested, and the respondents gave comments about the comprehensibility of the survey. The respondents are left with their assumptions on the detail of implementation and the objectives of some attributes (e.g., “number of free engaging hackathon events”, “number of free engaging data skill training”), and the impact of data leakage. For the final survey, descriptions of attributes purpose and examples of the implementation is provided.

Furthermore, the survey introduction uses several leading questions about respondents familiarity with open education data (open data portal that they have visited, services used, events attended) instead of long narratives. Introducing the open education data context using leading questions is chosen to prevent the respondents from skipping the introduction information. We realize that many respondents may not be familiar with the open education data topic and missing the context explanation will decrease the validity of their responses. The description is written in neutral wording to avoid anchoring effect in which respondent’s decision making are affected by the initial information provided.

Interpretation of citizen stated choice experiment (CSCE) result

The complexity of open data policy requires us to design abstraction of open education data policy in the form of ‘hypothetical situations’ and the respondents are asked to make their choice based on this abstraction.

In attributes selection, we can use more tangible attributes (number of datasets published, number of incidents occurred). However, we realized that would not mean anything if the respondents do not personally relate with the benefits from the selected attributes (quantity of data). What are the benefits of having more datasets if they cannot use it?

Therefore, in the design of choice situations and attributes selection we choose attributes that the respondents can personally valuate such as “risk of your personal data exposed to the public”, “mode of information presentation”, “number of free citizen data skill training events”, and “number of free engaging hackathon events”.

Especially for the “mode of information presentation” attribute, we define three categorical attribute levels which are presenting data in original form (as similar as possible with the source), in static and dynamic figure, and presenting information as a service. These attribute and attribute levels are based on the existing implementation of open education data policy and accompanied by examples of implementation as well. It is aimed to improve respondents perception of the survey validity and realism.

This study has several limitations, and the result should be used as the basis for further academic discussion rather than conclusive valuation of the attributes.

First, the target respondents for this study is limited to Dutch higher education students, and the content of the survey is designed to fit their context. Therefore, the result of this study is based on the preferences of Dutch higher education students in their role as a citizen. The result is based on 531 collected observations from 59 respondents (each of the respondent complete nine choice situations). Second, the possibility of ‘hypothetical bias’ due to the unfamiliarity of respondents to participation and engagement event attribute should be considered. It may skew the result because the respondent chooses attributes that are more familiar to them (mode of info presentation and data protection).

As in another validation study which employs a stated preference method, more empirical research is needed to validate the result. Johnston et al. (2017) said that,

> *“Assessment of the validity of any study or valuation method should consider the weight of the available evidence and should not depend on the outcome of a single test or investigation. Results of specific individual tests should not be considered as a prima facie justification for determining validity. Validity assessment should include study-specific design and analysis procedures and outcomes, as well as consideration of knowledge from the body of preceding research.”*

Academic contribution

Most of the previous OGD study uses qualitative methods (deep interview, desk research) to investigate the open data policy ecosystem and the quantitative approach are limited to open data portal assessments. Other than that, OGD field of study is dominated by studies based on data providers/government perspective. This study attempts to explore open education data policy from citizens (data users) perspective and empirically measure their preferences for open education data attributes.

The stated choice experiment has been widely used in the transport, health, and environmental valuation studies. This study is the first attempt to extend the utilization of stated preference (SP) method for the open government data domain. In this study, citizens preferences of open education data policy are empirically valuated using a variant of SP method called citizen stated choice experiment (CSCE).

From this study, we learn about the needs to adapt the design based on the context of the study. Many respondents are unfamiliar with open data policy context compared to transport, health, and environmental domain. Therefore, in the implementation of CSCE, we choose to provide more explanation for the context with the risk of ‘anchoring effect’ in order to improve respondents perception of the survey validity and realism. Otherwise, the respondents are left with their assumptions for the choice situation which may affect the validity of the responses. They might choose attributes that are more familiar to them and neglect less familiar attributes in their judgment.

7.11. CoSEM perspective

This research approaches the complex problem of open education data policy design. In the current conditions, policymakers implement open education data policy based on the established ‘best practice’ that they see from other countries or government agencies. However, this tendency to replicate the ‘best practice’ make policymakers ignore the unique policy context in which they operate. For example, the objectives for open education data or open geospatial data will be different and the potential end users of the datasets as well. There is no one size fits all policy, but in practice, policymakers implement a similar policy and evaluation framework. Policymakers focus on publishing as many datasets as possible and organizing one or two hackathons every year to engage the citizens. The development of open data policy from data provider perspective ignores the other side of open data (end users). Do the open data policy benefit end users and achieve the desired public values (transparency, accountability, and economic growth)?

These conditions show policymakers lack of insight into the citizen preferences of open data policy attributes. Open government data literature shows a need to balance data stewardship and usefulness capability of open government data program. There is a limited study that estimates the usefulness capability of the Dutch open education data policy. Therefore, this study attempts to use citizen stated choice experiments (CSCE) method to measure the citizens preference of open education data policy empirically.

The citizen stated choice experiments (CSCE) is a variant of the discrete choice experiment (DCE) which is learned in “Statistical Analysis of Choice Behavior" course of CoSEM study. In the creation of CSCE for open education data policy, it is important to understand the context of OGD and the existing body of knowledge. CoSEM I&C track “Integrated Design of I & C Architectures” course provides the basis for the investigation, especially the topic of G2C (Government to Citizens) interactions.

In this research, the CoSEM perspective helps us to investigate the open education data policy from both data providers and data users perspective. The policy context exploration is similar to system analysis in the systems engineering approach in which the policy objectives, organizational context and existing policy implementation are investigated. The collected information is used to design a survey that clearly explains the policy context for the respondents so that they can make an informed choice regarding their preferences for open education data policy. Other than that, feedback from the respondents are also important to improve the realism and validity of the survey. A pilot test is conducted with a limited number of respondents, and the feedbacks are used to improve the final survey. CoSEM perspective enables us to investigate the problem from both perspectives and synthesize the result into the design of citizen stated choice experiment (CSCE) to analyze citizens preferences for a Dutch open education data policy.

The result of this method can be used to complement the existing evaluation and development of open data policy which uses the data providers perspective. Using the CSCE method presented in this study, policymakers can obtain a better insight into the risk and benefits of opening data from both data providers and data users perspective. The result can also be used to justify their choices in the decision-making process, why they choose one attributes over the others.

This research not only results in the citizens valuation of open education data policy attributes but also the approach to design similar CSCE in the different open data context. The design of CSCE in this research can be modified by policymakers for other open government data context such as (geospatial data, assets inventory data, spending data, open data portal features) or even same context (education data) with different target respondents (parents and primary/secondary education students).

Bibliography

Achtnicht, M. (2011). Do environmental benefits matter? Evidence from a choice experiment among house owners in Germany. *Ecological Economics*, *70*(11), 2191–2200. https://doi.org/10.1016/j.ecolecon.2011.06.026

Afful-Dadzie, E., & Afful-Dadzie, A. (2017). Open Government Data in Africa: A preference elicitation analysis of media practitioners. *Government Information Quarterly*, *34*(2), 244–255. https://doi.org/10.1016/j.giq.2017.02.005

Algemene Rekenkamer. (2014). *Trend Report Open Data*. Retrieved from https://english.rekenkamer.nl/binaries/rekenkamer-english/documents/reports/2014/03/27/open-data-trend-report/Trend+Report+Open+Data+2014.pdf

Algemene Rekenkamer. (2016). *Trendrapport open data 2016*. Retrieved from http://www.rekenkamer.nl/Publicaties/Onderzoeksrapporten/Introducties/2015/03/Trendrapport_open_data_2015

Arrow, K., Solow, R., Portney, P., Leamer, E., Radner, R., & Schuman, H. (1993). *Report of the NOAA panel on Contingent Valuation*. *Federal Register* (Vol. 58).

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Bierlaire, M. (2016). PythonBiogeme : a short introduction. *Report TRANSP-OR 160706, Series on Biogeme. Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Switzerland.*

Carson, R. T., & Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics*, *37*(1), 181–210. https://doi.org/10.1007/s10640-007-9124-5

Charalabidis, Y., Alexopoulos, C., & Loukis, E. (2016). A taxonomy of open government data research areas and topics. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 41–63. https://doi.org/10.1080/10919392.2015.1124720

Chatfield, A. T., & Reddick, C. G. (2017). A longitudinal cross-sector analysis of open data portal service capability: The case of Australian local governments. *Government Information Quarterly*, *34*(2), 231–243. https://doi.org/10.1016/j.giq.2017.02.004

Cheraghi-Sohi, S., Hole, A. R., Mead, N., McDonald, R., Whalley, D., Bower, P., & Roland, M. (2008). What patients want from primary care consultations: A discrete choice experiment to identify patients' priorities. *Annals of Family Medicine*, *6*(2), 107–115. https://doi.org/10.1370/afm.816

Cook, D., Davídsdóttir, B., & Kristófersson, D. M. (2016). Energy projects in Iceland - Advancing the case for the use of economic valuation techniques to evaluate environmental impacts. *Energy Policy*, *94*, 104–113. https://doi.org/10.1016/j.enpol.2016.03.044

data.overheid.nl. (n.d.). Score van Nederland in benchmarks. Retrieved May 14, 2018, from https://data.overheid.nl/score-van-nederland-benchmarks data.overheid.nl. (2018). Open Data Policy. Retrieved April 20, 2018, from https://data.overheid.nl/open-data-beleid

Dawes, S. S. (2010). Stewardship and usefulness: Policy principles for information-based transparency. *Government Information Quarterly*, *27*(4), 377–383. https://doi.org/10.1016/j.giq.2010.07.001

Dawes, S. S., Vidiasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, *33*(1), 15–27. https://doi.org/10.1016/j.giq.2016.01.003

DeShazo, J. R., & Fermo, G. (2002). Designing choice sets for stated preference methods: The effects of complexity on choice consistency. *Journal of Environmental Economics and Management*, *44*(1), 123–143. https://doi.org/10.1006/jeem.2001.1199

European Commission. (2003). Directive 2003/98/EC of the European Parliament and of the council of 17 November 2003 on the re-use of public sector information. Retrieved March 15, 2018, from https://ec.europa.eu/digital-single-market/overview-2003-psi-directive

European Commission. (2011). Digital Agenda: Commission’s Open Data Strategy, Questions & answers. Retrieved March 15, 2018, from http://europa.eu/rapid/press-release_MEMO-11-891_en.htm?locale=en

Hall, J., Viney, R., Haas, M., & Louviere, J. (2004). Using stated preference discrete choice modeling to evaluate health care programs. *Journal of Business Research*, *57*(9), 1026–1032. https://doi.org/10.1016/S0148-2963(02)00352-1

Hanne Obbink. (2012, October 11). Er zijn wél slechte scholen | TROUW. Retrieved from https://www.trouw.nl/home/er-zijn-wel-slechte-scholen~a56a8c06/

Hauber, A. B., González, J. M., Groothuis-Oudshoorn, C. G. M., Prior, T., Marshall, D. A., Cunningham, C., … Bridges, J. F. P. (2016). Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force. *Value in Health*, *19*(4), 300–315. https://doi.org/10.1016/j.jval.2016.04.004

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Johnston, R. J., Boyle, K. J., Adamowicz, W. (Vic), Bennett, J., Brouwer, R., Cameron, T. A., … Vossler, C. A. (2017). Contemporary Guidance for Stated Preference Studies. *Journal of the Association of Environmental and Resource Economists*, *4*(2), 319–405. https://doi.org/10.1086/691697

Kjær, T., & Gyrd-Hansen, D. (2008). Preference heterogeneity and choice of cardiac rehabilitation program: Results from a discrete choice experiment. *Health Policy*, *85*(1), 124–132. https://doi.org/10.1016/j.healthpol.2007.07.002

Kuhfeld, W. F. (2010). Marketing research methods in SAS experimental design, choice, conjoint, and graphical techniques. *… Graphical Techniques. Cary, NC, SAS-Institute TS-722*, 1–1309. Retrieved from http://www.soc.iastate.edu/Sapp/soc512Kuhfeld.pdf%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.163.8176

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*(2), 132–157. https://doi.org/10.1086/259131

Lee, G., & Kwak, Y. H. (2012). An Open Government Maturity Model for social media-based public engagement. *Government Information Quarterly*, *29*(4), 492–503. https://doi.org/10.1016/j.giq.2012.06.001

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, *32*(3), 323–332. https://doi.org/10.1016/j.giq.2015.05.006

Mangham, L. J., Hanson, K., & McPake, B. (2009). How to do (or not to do)...Designing a discrete choice experiment for application in a low-income country. *Health Policy and Planning*, *24*(2), 151–158. https://doi.org/10.1093/heapol/czn047

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (pp. 105–142). https://doi.org/10.1108/eb028592

Ministry of Education, C. and S. (2015). *Transparantie in het funderend onderwijs*. Retrieved from https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/kamerstukken/2015/10/26/kamerbrief-over-transparantie-in-het-onderwijs/kamerbrief-over-transparantie-in-het-onderwijs.pdf

Ministry of the Interior and Kingdom Relations. (2015). *Kamerbrief over nationale open data agenda 2016 (NODA)*. Retrieved from https://www.rijksoverheid.nl/documenten/kamerstukken/2015/11/30/kamerbrief-over-nationale-open-data-agenda-2016-noda

Ministry of the Interior and Kingdom Relations. (2017). *Netherlands Mid-Term Self-Assessment Report National Action Plan Open Government 2016-2018*. Retrieved from https://www.opengovpartnership.org/sites/default/files/Netherlands_Mid-term_Self-Assessment-Report_2016-2018_EN.pdf

Mouter, N., & Chorus, C. (2016). Value of time – A citizen perspective. *Transportation Research Part A: Policy and Practice*, *91*, 317–329. https://doi.org/10.1016/j.tra.2016.02.014

Mouter, N., van Cranenburgh, S., & van Wee, B. (2017a). An empirical assessment of Dutch citizens' preferences for spatial equality in the context of a national transport investment plan. *Journal of Transport Geography*, *60*, 217–230. https://doi.org/10.1016/j.jtrangeo.2017.03.011

Mouter, N., van Cranenburgh, S., & van Wee, B. (2017b). Do individuals have different preferences as consumer and citizen? The trade-off between travel time and safety. *Transportation Research Part A: Policy and Practice*, *106*(September 2016), 333–349. https://doi.org/10.1016/j.tra.2017.10.003

Obama, B. (2009). Open government directive. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

Obama, B. (2012). Digital Government: Building a 21st Century Platform to Better Serve the American People. Retrieved March 15, 2018, from https://obamawhitehouse.archives.gov/sites/default/files/omb/egov/digital-government/digital-government.html

OCW. (2017). *Rijksbegroting 2018*. https://doi.org/ISSN 09217371

OCW. (2018). *Rijksjaarverslag 2017 VIII Onderwijs, Cultuur en Wetenschap*. https://doi.org/ISSN 09217371

open-overheid.nl. (2018). Brief history of the Learning and Expertise Point. Retrieved April 21, 2018, from https://www.open-overheid.nl/open-overheid/the-making-of-open-overheid/

openstate.eu. (2016). Dutch Ministry of Education launches open education API – Open State Foundation. Retrieved April 22, 2018, from https://openstate.eu/en/2016/11/dutch-ministry-of-education-launches-open-education-api/

openstate.eu. (2018). Amsterdam kicks off with a hackathon series about education – Open State Foundation. Retrieved April 22, 2018, from https://openstate.eu/en/2018/02/amsterdam-kicks-off-with-a-hackathon-series-about-education/

Petychakis, M., Vasileiou, O., Georgis, C., Mouzakitis, S., & Psarras, J. (2014). A state-of-the-art analysis of the current public data landscape from a functional, semantic and technical perspective. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*(2), 34–47. https://doi.org/10.4067/S0718-18762014000200004

Reggi, L., & Ricci, C. A. (2011). Information strategies for open government in Europe: EU regions opening up the data on structural funds. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6846 LNCS, pp. 173–184). https://doi.org/10.1007/978-3-642-22878-0_15

Rijksoverheid. (2013a). *Open Government Action Plan*. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2013/09/01/actieplan-open-overheid

Rijksoverheid. (2013b). *Vision Open Government*. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2013/09/01/visie-open-overheid

Rijksoverheid. (2016). Data-Expedition report Ministry of Education, Culture and Science Rijksoverheid.nl. Retrieved April 22, 2018, from https://www.rijksoverheid.nl/ministeries/ministerie-van-onderwijs-cultuur-en-wetenschap/evenementen/onderwijsdata-onder-de-loep/data-expeditie

Rijksoverheid. (2018a). Open Government. Retrieved April 20, 2018, from https://www.rijksoverheid.nl/onderwerpen/digitale-overheid/open-overheid

Rijksoverheid. (2018b). Openheid over kwaliteit onderwijs. Retrieved April 22, 2018, from https://www.rijksoverheid.nl/onderwerpen/openheid-over-kwaliteit-onderwijs

Rubin, G., Bate, A., & George, A. (2006). Preferences for access to the GP: a discrete choice experiment. *British Journal of General Practice*, *56*, 743–748.

Ryan, M. (2004). Discrete choice experiments in health care. *BMJ (Clinical Research Ed.)*, *328*(7436), 360–1. https://doi.org/10.1136/bmj.328.7436.360

Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity*, *22*(1), 1–

24. https://doi.org/10.3233/IP-160012

Sayogo, D. S., Pardo, T. A., & Cook, M. (2014). A framework for benchmarking open government data efforts. *Proceedings of the Annual Hawaii International Conference on System Sciences*, (May 2010), 1896–1905. https://doi.org/10.1109/HICSS.2014.240

scholenopdekaart.nl. (2018). Vind en vergelijk scholen bij jou in de buurt. Retrieved April 22, 2018, from https://www.scholenopdekaart.nl/

Sieber, R. E., & Johnson, P. A. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly*, *32*(3), 308–315. https://doi.org/10.1016/j.giq.2015.05.003

Susha, I., Zuiderwijk, A., Janssen, M., & Grönlund, Å. (2015). Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned. *Social Science Computer Review*, *33*(5), 613–630. https://doi.org/10.1177/0894439314560852

Thorsby, J., Stowers, G. N. L., Wolslegel, K., & Tumbuan, E. (2017). Understanding the content and features of open data portals in American cities. *Government Information Quarterly*, *34*(1), 53–61. https://doi.org/10.1016/j.giq.2016.07.001

Tim Berners-Lee. (n.d.). 5-star Open Data. Retrieved May 5, 2018, from http://5stardata.info/en/

Titah, J. H. R. (2017). Conceptualizing citizen participation in open data use at the city level. *Transforming Government: People, Process and Policy*, *11*(1), 99–118. https://doi.org/10.1108/TG-12-2015-0053

Train, K. E. (2003). Discrete Choice Methods with Simulation. *Cambridge University Press*, 1–388. https://doi.org/10.1017/CBO9780511753930

Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*, *NO.22*(22), 61. https://doi.org/10.1787/5k46bj4f03s7-en

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337. https://doi.org/10.1016/j.giq.2016.02.001

Walker, J. L., Wang, Y., Thorhauge, M., & Ben-Akiva, M. (2018). D-efficient or deficient? A robustness analysis of stated choice experimental designs. *Theory and Decision*, *84*(2), 215–238. https://doi.org/10.1007/s11238-017-9647-3

Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., & Dwivedi, Y. K. (2017). Open data and its usability: an empirical view from the Citizen's perspective. *Information Systems Frontiers*, *19*(2), 285–300. https://doi.org/10.1007/s10796-016-9679-1

Welle Donker, F., & van Loenen, B. (2017). How to assess the success of the open data ecosystem? *International Journal of Digital Earth*, *10*(3), 284–306. https://doi.org/10.1080/17538947.2016.1224938

Welle Donker, F., van Loenen, B., & Korthals Altes, W. (2017). *Maatschappelijke kosten-batenanalyse open data*.

Westra, E., & Poel, R. van der. (2017). *De websites met statistieken over de stelsels van het ministerie van OCW*. Amsterdam. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2017/07/01/de-websites-met-statistieken-over-de-stelsels-van-het-ministerie-van-ocw---verslag-van-een-onderzoek

Zuiderwijk-van Eijk, A. M. G., & Janssen, M. F. W. H. A. (2015). Participation and Data Quality in Open Data use: Open Data Infrastructures Evaluated. *Proceedings of The15th European Conference on E-Government, Portsmouth, UK, 18-19 June 2015; Authors Version*. Retrieved from https://repository.tudelft.nl/islandora/object/uuid:c3e2530d-eaa2-409b-a700-b7107db7e159?collection=research

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, *31*(1), 17–29. https://doi.org/10.1016/j.giq.2013.04.003

Zuiderwijk, A., Shinde, R., & Janssen, M. (2018). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results? *International Review of Administrative Sciences*. https://doi.org/10.1177/0020852317739115

Appendix

A. Survey design

Ngene design generation:

The syntax that is required to generate the choice sets with Ngene is shown in the textbox below.

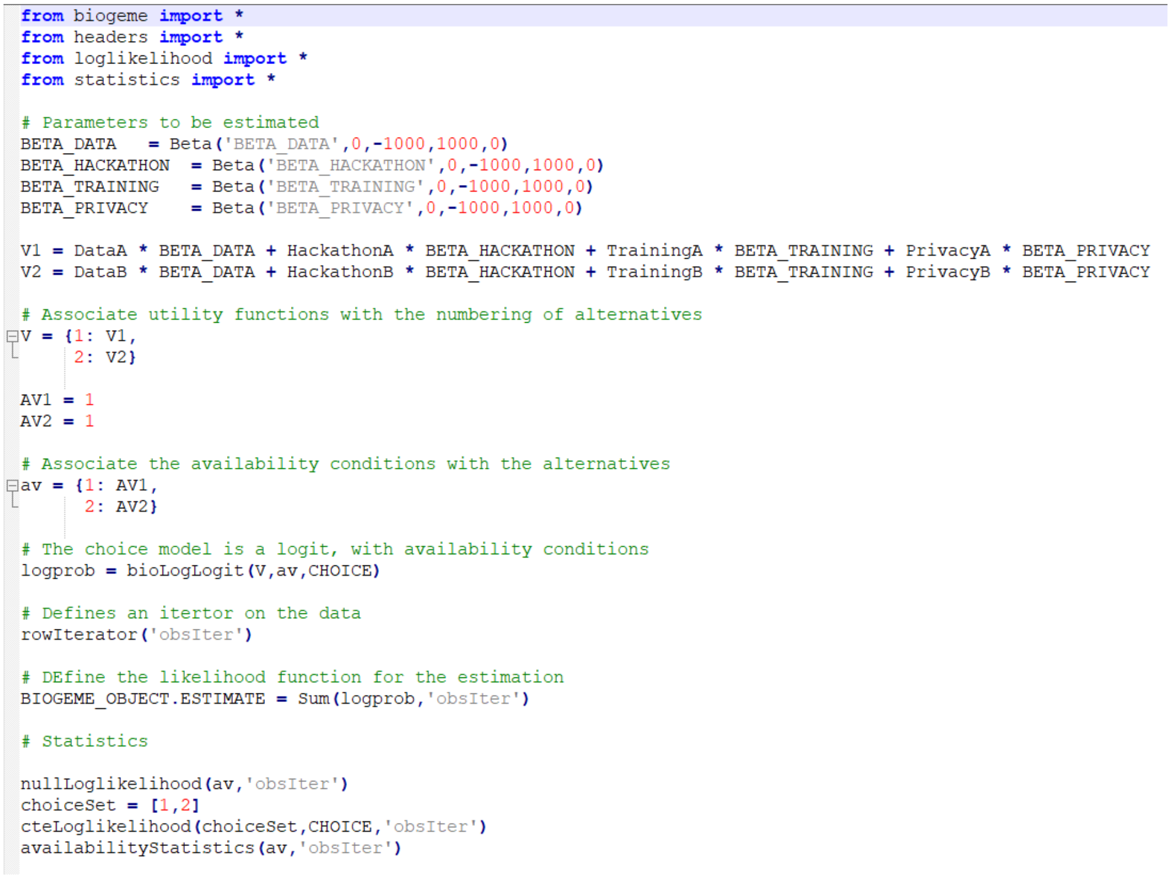
```
design
; alts = alt1, alt2
; rows = 9
; orth = seq
; model:
U(alt1) = b1 * Wdata[0,1,2] + b2 * Whackathon[0.5,1,2] + b3 *
Wtraining[1,2,3] + b4 * Wprivacy[0,1,2] /
U(alt2) = b1 * Wdata + b2 * Whackathon + b3 * training + b4 * Wprivacy
$
```

The code defines four attributes, 9 choice sets, a sequential orthogonal design and the two utility functions for two unlabeled alternatives.

Model file:

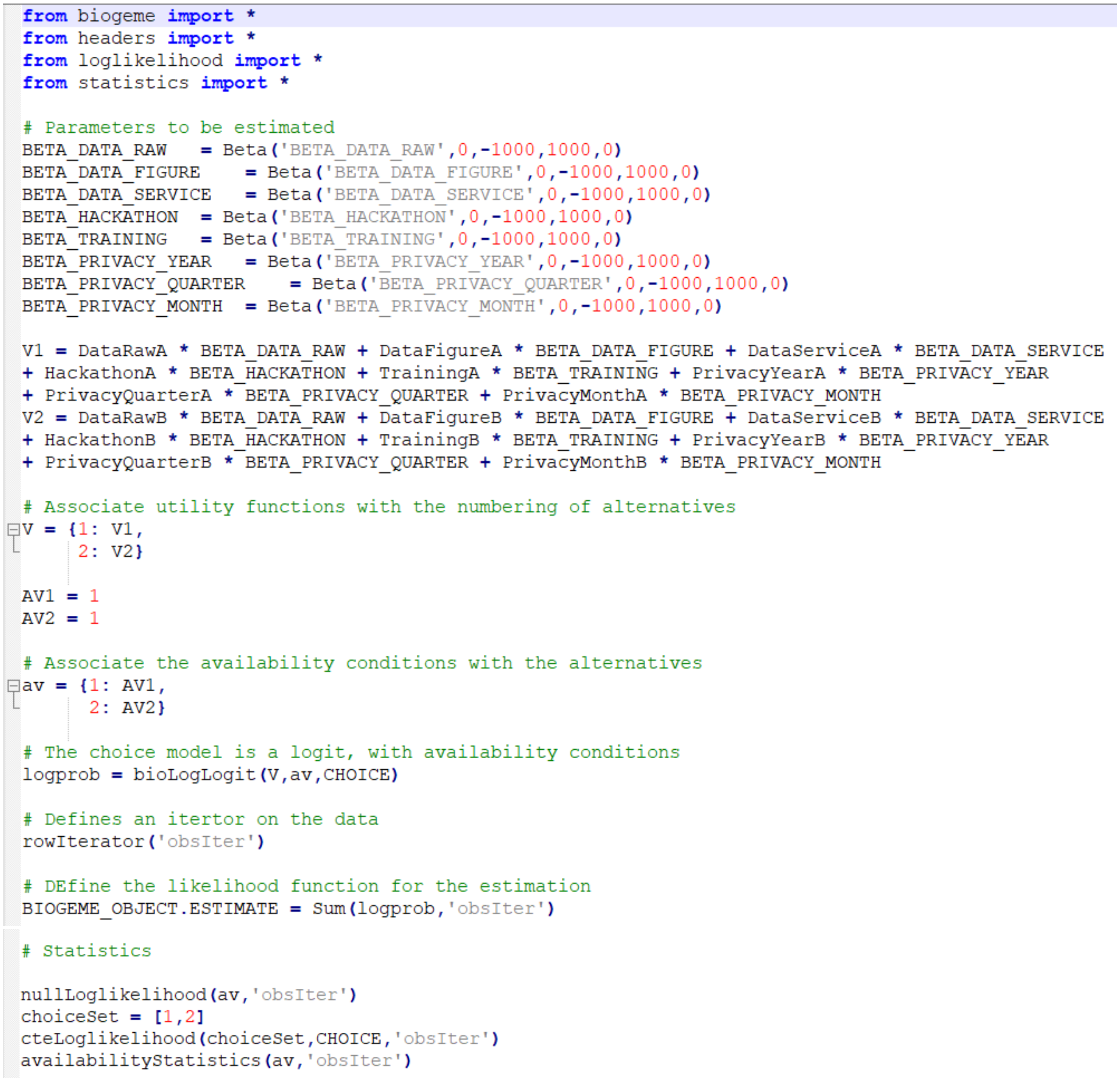
The model file specification is depicted in the code below:

1. Initial model to calculate the parameters utility



```
from biogeme import *
from headers import *
from loglikelihood import *
from statistics import *

# Parameters to be estimated
BETA_DATA   = Beta('BETA_DATA',0,-1000,1000,0)
BETA_HACKATHON  = Beta('BETA_HACKATHON',0,-1000,1000,0)
BETA_TRAINING   = Beta('BETA_TRAINING',0,-1000,1000,0)
BETA_PRIVACY    = Beta('BETA_PRIVACY',0,-1000,1000,0)

V1 = DataA * BETA_DATA + HackathonA * BETA_HACKATHON + TrainingA * BETA_TRAINING + PrivacyA * BETA_PRIVACY
V2 = DataB * BETA_DATA + HackathonB * BETA_HACKATHON + TrainingB * BETA_TRAINING + PrivacyB * BETA_PRIVACY

# Associate utility functions with the numbering of alternatives
V = {1: V1,
     2: V2}

AV1 = 1
AV2 = 1

# Associate the availability conditions with the alternatives
av = {1: AV1,
      2: AV2}

# The choice model is a logit, with availability conditions
logprob = bioLogLogit(V,av,CHOICE)

# Defines an itertor on the data
rowIterator('obsIter')

# DEfine the likelihood function for the estimation
BIOGEME_OBJECT.ESTIMATE = Sum(logprob,'obsIter')

# Statistics

nullLoglikelihood(av,'obsIter')
choiceSet = [1,2]
cteLoglikelihood(choiceSet,CHOICE,'obsIter')
availabilityStatistics(av,'obsIter')
```

2. Model to check parameters linearity



```
from biogeme import *
from headers import *
from loglikelihood import *
from statistics import *

# Parameters to be estimated
BETA_DATA_RAW   = Beta('BETA_DATA_RAW',0,-1000,1000,0)
BETA_DATA_FIGURE    = Beta('BETA_DATA_FIGURE',0,-1000,1000,0)
BETA_DATA_SERVICE   = Beta('BETA_DATA_SERVICE',0,-1000,1000,0)
BETA_HACKATHON  = Beta('BETA_HACKATHON',0,-1000,1000,0)
BETA_TRAINING   = Beta('BETA_TRAINING',0,-1000,1000,0)
BETA_PRIVACY_YEAR   = Beta('BETA_PRIVACY_YEAR',0,-1000,1000,0)
BETA_PRIVACY_QUARTER    = Beta('BETA_PRIVACY_QUARTER',0,-1000,1000,0)
BETA_PRIVACY_MONTH  = Beta('BETA_PRIVACY_MONTH',0,-1000,1000,0)

V1 = DataRawA * BETA_DATA_RAW + DataFigureA * BETA_DATA_FIGURE + DataServiceA * BETA_DATA_SERVICE
+ HackathonA * BETA_HACKATHON + TrainingA * BETA_TRAINING + PrivacyYearA * BETA_PRIVACY_YEAR
+ PrivacyQuarterA * BETA_PRIVACY_QUARTER + PrivacyMonthA * BETA_PRIVACY_MONTH
V2 = DataRawB * BETA_DATA_RAW + DataFigureB * BETA_DATA_FIGURE + DataServiceB * BETA_DATA_SERVICE
+ HackathonB * BETA_HACKATHON + TrainingB * BETA_TRAINING + PrivacyYearB * BETA_PRIVACY_YEAR
+ PrivacyQuarterB * BETA_PRIVACY_QUARTER + PrivacyMonthB * BETA_PRIVACY_MONTH

# Associate utility functions with the numbering of alternatives
V = {1: V1,
     2: V2}

AV1 = 1
AV2 = 1

# Associate the availability conditions with the alternatives
av = {1: AV1,
      2: AV2}

# The choice model is a logit, with availability conditions
logprob = bioLogLogit(V,av,CHOICE)

# Defines an itertor on the data
rowIterator('obsIter')

# DEfine the likelihood function for the estimation
BIOGEME_OBJECT.ESTIMATE = Sum(logprob,'obsIter')

# Statistics

nullLoglikelihood(av,'obsIter')
choiceSet = [1,2]
cteLoglikelihood(choiceSet,CHOICE,'obsIter')
availabilityStatistics(av,'obsIter')
```

B. Pilot test feedback

Respondent 1	• Most important attribute: Mode of information presentation • Least important attribute: Number of free engaging hackathon events Comments: • Provide information on the effect of hackathon and data training, what the hackathon and data training do? • Make the result of hackathon and data training tangible in the mind of respondents
Respondent 2	• Most important attribute: Risk of your personal data exposed to the public • Least important attribute: Number of free engaging hackathon events

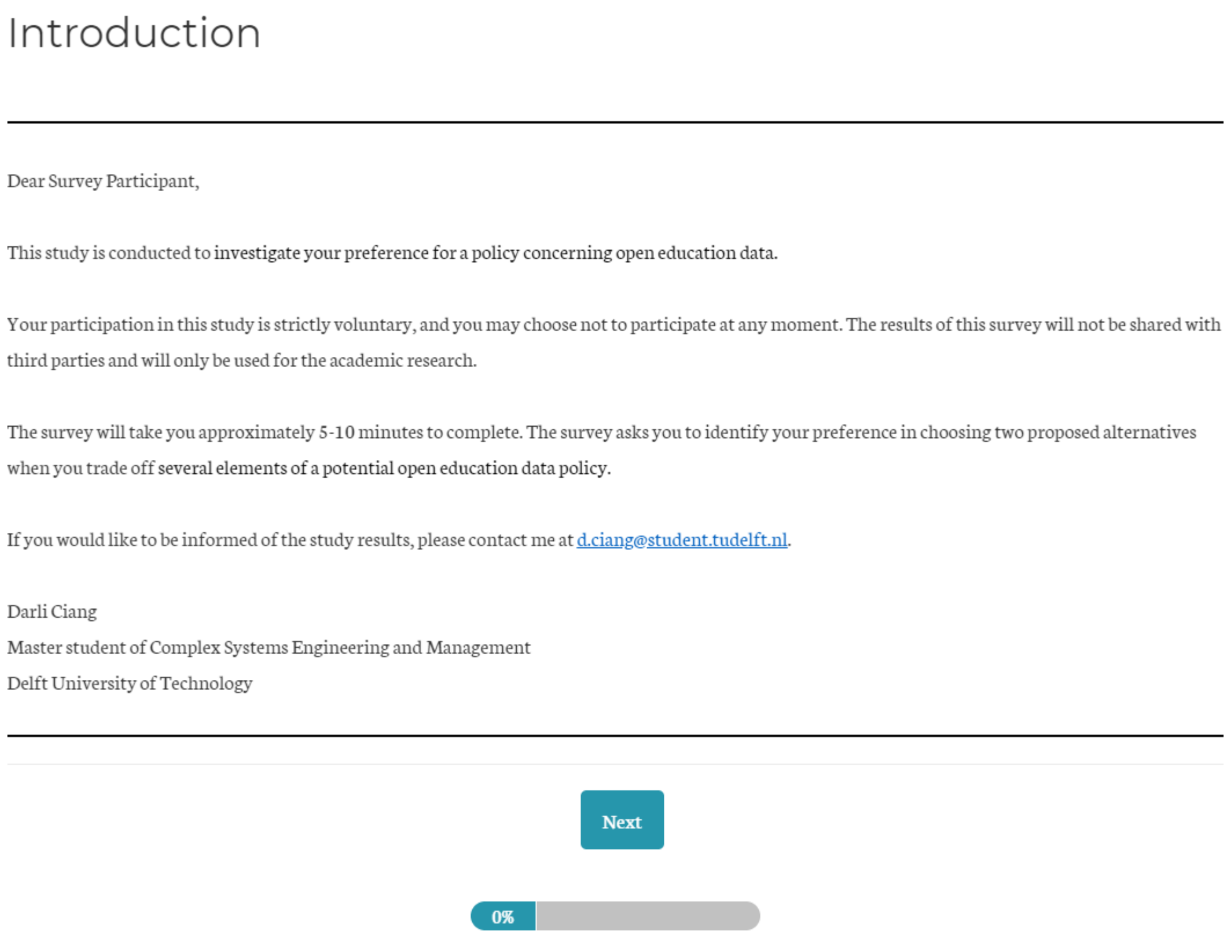
	Comments: I think the privacy aspect is most important for most people and the other aspects are less relevant. Since I'm not really into data science, I find it difficult to make the trade-offs and I merely looked at the data leakage aspect
Respondent 3	• Most important attribute: Risk of your personal data exposed to the public • Least important attribute: Mode of information presentation Comments: Have difficulty to choose between 4 attributes
Respondent 4	• Most important attribute: Mode of information presentation • Least important attribute: Number of free citizen data skill training events Comments: • Provide better description about the attributes level/context. People with technical and non-technical background can have a different interpretation if the description is not clear. • To what extent the hackathon is conducted? 1 hackathon with 100 participants is different from 2 hackathons with 30 participants • Data skill training and hackathon may become less important because of the respondents' educational background
Respondent 5	• Most important attribute: Risk of your personal data exposed to the public • Least important attribute: Mode of information presentation Comments: I consider it hard for me to provide you with recommendations to improve this questionnaire. Since I consider this topic a bit abstract. However, in the first place I did not properly understand the relation between data training, hackathon events and open education policy. Therefore, I could not assess if for example a lot of hackathon events is a beneficial in a certain policy. same applies for the data skill training events. Therefore, I should recommend to explain these aspects better in your questionnaire.

	In addition, I did not fully understand your 7 points scale for question 5: why didn't you apply a 5 points scale or a ten points scale.
Respondent 6	• Most important attribute: Risk of your personal data exposed to the public • Least important attribute: Number of free engaging hackathon events Comments: The choice situations were quite a bit of reading... If possible, I would reduce the amount of text so that comparing the alternatives is easier. An example might be to just state "2 per year" in the case of free hackathon / data skill training events (as that is already specified in the left column). On a more general note, why did you only include these events as possibilities? Would it not be far more convenient for a lot of people to create, for instance, an online learning environment for people to get acquainted with open data? (If there is a good motivation for it, neglect this comment). It feels to me as if the choice sets are fairly limited at this point, although I do understand that it might get very complicated and time-consuming for the respondents if you were to expand on it.
Respondent 7	• Most important attribute: Mode of information presentation • Least important attribute: Number of free engaging hackathon events Comments: • Is it anonymous for the participants? No third-party access, but what happens when the paper is published? • You define what is open data to the participant, but not what open educational data? Is there a difference and how? • Question three, are there more options? And is it possible to increase the size of the pictures for a better view? Look out for mobile users with a smaller screen? • Question 4 try to find as many options as possible, people are lazy and are not going to find the websites themselves • Question 5 define difference open educational data and personal data. Is there personal data in the open educational data? • Define free engaging hackathon, nearly all companies and governments organize hackathons and are all very different from a business optimization to a real hackathon whereby students try to hack a system and gain important information. What is the purpose? • Question 6 change the number words to real numbers. Makes it easier to compare. For example, one -> 1x • Explain what are the free engaging hackathon, why, how, where, when and for who?

	• Explain what are the skill trainings for why, how, where, when and for who? • Time is a lot shorter. It took me 8 minutes.
Respondent 8	• Most important attribute: Risk of your personal data exposed to the public • Least important attribute: Number of free engaging hackathon events Comments: • Q5: you could specify what is meant by personal data (your bank account? Your education level?) • As for the questionnaire overall: I think this might be more relevant for a group that already works with the data/topic, I feel like a lot of citizens won't really have an opinion on the matter (unless that is obviously something you plan to measure) • For me personally, stuff like hackathons and data training are not very interesting. So, it's not necessary the information, it's more that the topic does not relate to me personally
Respondent 9	• Most important attribute: Risk of your personal data exposed to the public • Least important attribute: Number of free engaging hackathon events Comments: Opening • You could add the purpose of the questionnaire, to assist you in graduation, maybe people are more obliged to fill it in if they do it for you instead of the government or the sake of open data • You request some private data at the end, but no identifiable information, maybe state this in the opening Questions page 1 • As you mentioned, you are hoping to target students that are familiar with open data, then the introduction of the concept stands. Yet, I think many students are not familiar with the concept but do use it simply because they need data for their thesis, like we do. I would therefore advice to also included some potential use cases of open data so maybe people recognize that they are actually familiar. • Question 5 might need some more introduction as it is quite different from the other questions and might scare away a bit. Plus, is the data leak caused by opening up data or by other practices? Questions page 2 • Maybe in the overview instead of bullet points give them ABC, or option 1,2,3. This makes it clearer that people should pick one. It works like this too, but I only understood seeing the questions.

	Questions page 3 • Is there a specific reason you categorize age this way? • Majoring is not a term often used in Dutch, maybe use Specialization, also you will get many different answers here, you could provide categories.
Respondent 10	• Most important attribute: Mode of information presentation • Least important attribute: Number of free citizen data skill training events Comments: -

C. Final Survey



Introduction

Dear Survey Participant,

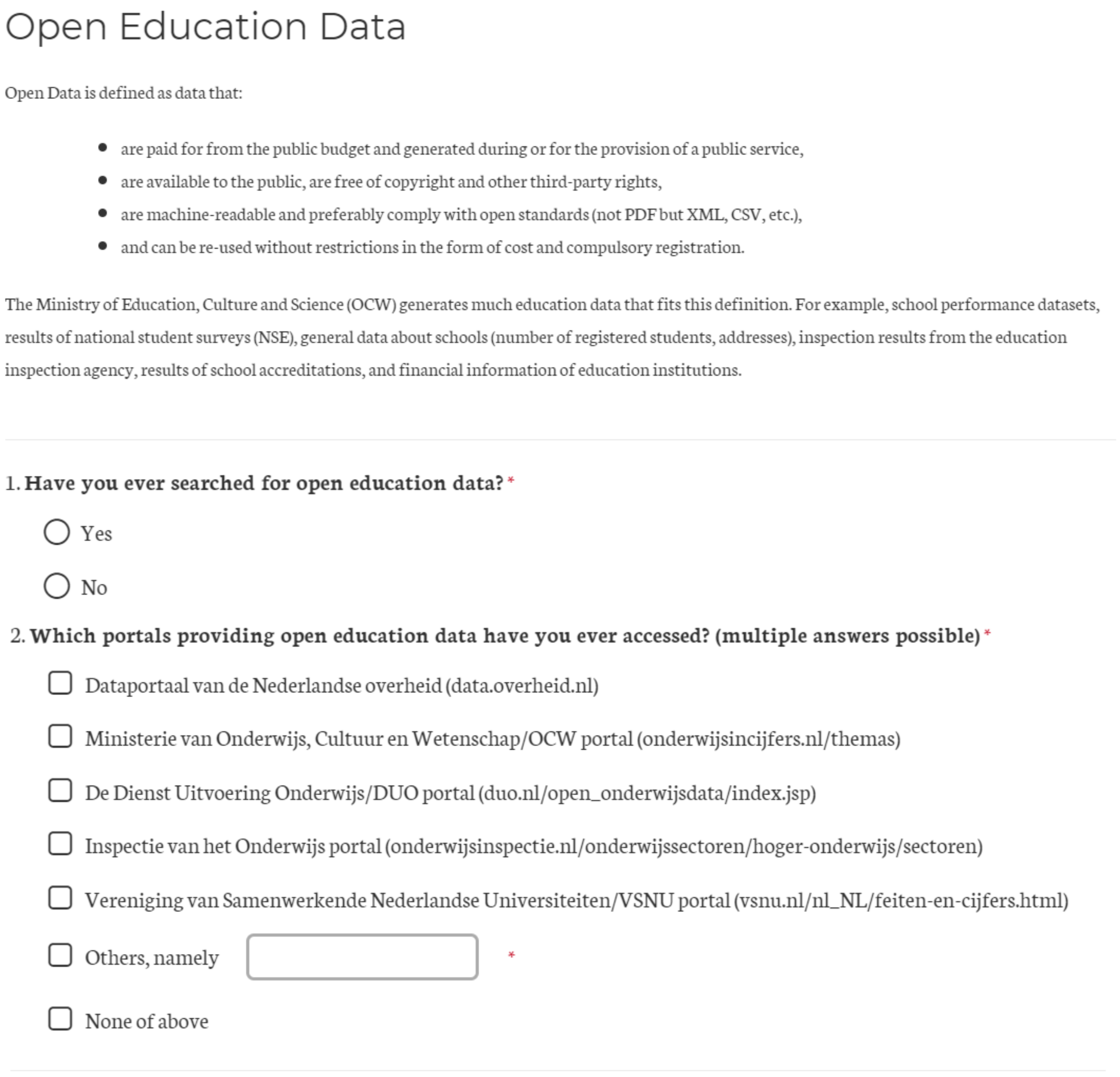
This study is conducted to investigate your preference for a policy concerning open education data.

Your participation in this study is strictly voluntary, and you may choose not to participate at any moment. The results of this survey will not be shared with third parties and will only be used for the academic research.

The survey will take you approximately 5-10 minutes to complete. The survey asks you to identify your preference in choosing two proposed alternatives when you trade off several elements of a potential open education data policy.

If you would like to be informed of the study results, please contact me at d.ciang@student.tudelft.nl.

Darli Ciang
Master student of Complex Systems Engineering and Management
Delft University of Technology



Open Education Data

Open Data is defined as data that:

- are paid for from the public budget and generated during or for the provision of a public service,
- are available to the public, are free of copyright and other third-party rights,
- are machine-readable and preferably comply with open standards (not PDF but XML, CSV, etc.),
- and can be re-used without restrictions in the form of cost and compulsory registration.

The Ministry of Education, Culture and Science (OCW) generates much education data that fits this definition. For example, school performance datasets, results of national student surveys (NSE), general data about schools (number of registered students, addresses), inspection results from the education inspection agency, results of school accreditations, and financial information of education institutions.

1. **Have you ever searched for open education data?** *

○ Yes

○ No

2. **Which portals providing open education data have you ever accessed? (multiple answers possible)** *

□ Dataportaal van de Nederlandse overheid (data.overheid.nl)

□ Ministerie van Onderwijs, Cultuur en Wetenschap/OCW portal (onderwijsincijfers.nl/themas)

□ De Dienst Uitvoering Onderwijs/DUO portal (duo.nl/open_onderwijsdata/index.jsp)

□ Inspectie van het Onderwijs portal (onderwijsinspectie.nl/onderwijssectoren/hoger-onderwijs/sectoren)

□ Vereniging van Samenwerkende Nederlandse Universiteiten/VSNU portal (vsnu.nl/nl_NL/feiten-en-cijfers.html)

□ Others, namely ______ *

□ None of above

3. **Which services created using open education data have you ever used? (multiple answers possible)**

Wegwijzers
Opleidingen
Open dagen
Instellingen
IN TWEE JAAR EEN ERKEND HBO-DIPLOMA?
Check alle associate degree-opleidingen
Zoek hier jouw opleiding...
MBO

studiekeuze123.nl

Scholen op de kaart
Vind en vergelijk scholen bij jou in de buurt
Ik zoek een
basisschool
middelbare school
Zoek op postcode, plaats of schoolnaam
Zoek
Gebruik mijn locatie

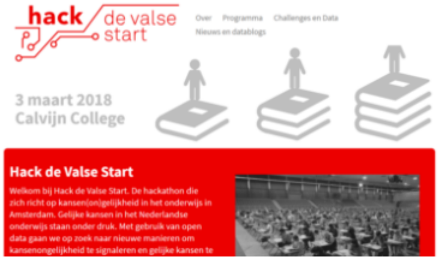
Wat biedt Scholen op de kaart?

Op deze site vind je veelzijdige informatie over basisscholen en middelbare scholen en kan je scholen vergelijken. Of je nou op het punt staat om een goede en geschikte school te kiezen, of als professional benieuwd bent hoe het staat met de kwaliteit van de scholen in je omgeving, voor iedereen biedt deze site interessante informatie over onderwijs en resultaten.

scholenopdekaart.nl *

- □ studiekeuze123.nl
- □ scholenopdekaart.nl
- □ Others, namely ______ *
- □ None of those

4. **Which open data events organized by the government have you ever participated in? (multiple answers possible)**



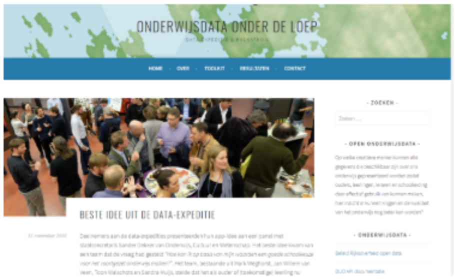
Over Programma Challenges en Data Nieuws en datablogs

3 maart 2018
Calvijn College

Hack de Valse Start

Welkom bij Hack de Valse Start. De hackathon die zich richt op kansen(on)gelijkheid in het onderwijs in Amsterdam. Gelijke kansen in het Nederlandse onderwijs staan onder druk. Met gebruik van open data gaan we op zoek naar nieuwe manieren om kansenongelijkheid te signaleren en gelijke kansen te

hackdevalsestart.nl



ONDERWIJSDATA ONDER DE LOEP

HOME · OVER · TOOLKIT · RESULTATEN · CONTACT

BESTE IDEE UIT DE DATA-EXPEDITIE

- ZOEKEN -

Zoeken ...

- OPEN ONDERWIJSDATA -

- ONDERWIJSDATA -

onderwijsdata.wordpress.com/blog/ *

□ The Hackathon ‘Hack de Valse Start’ organized by the Ministry of Education, Culture and Science in March 2018

□ The data exploration event ‘Onderwijsdata Onder de Loep’ organized by the Ministry of Education, Culture and Science in November 2016

□ Other(s), namely [ ] *

□ None of those

5. **On a scale from 1 to 5, to what extent are you concerned that the government will violate your privacy through the leakage of your personal data*?**

*The open education data is anonymized. The personal data leakage happens when a person can be identified from the combination of multiple anonymous open datasets.

*

	1	2	3	4	5	
Not concerned at all	○	○	○	○	○	**Extremely concerned**

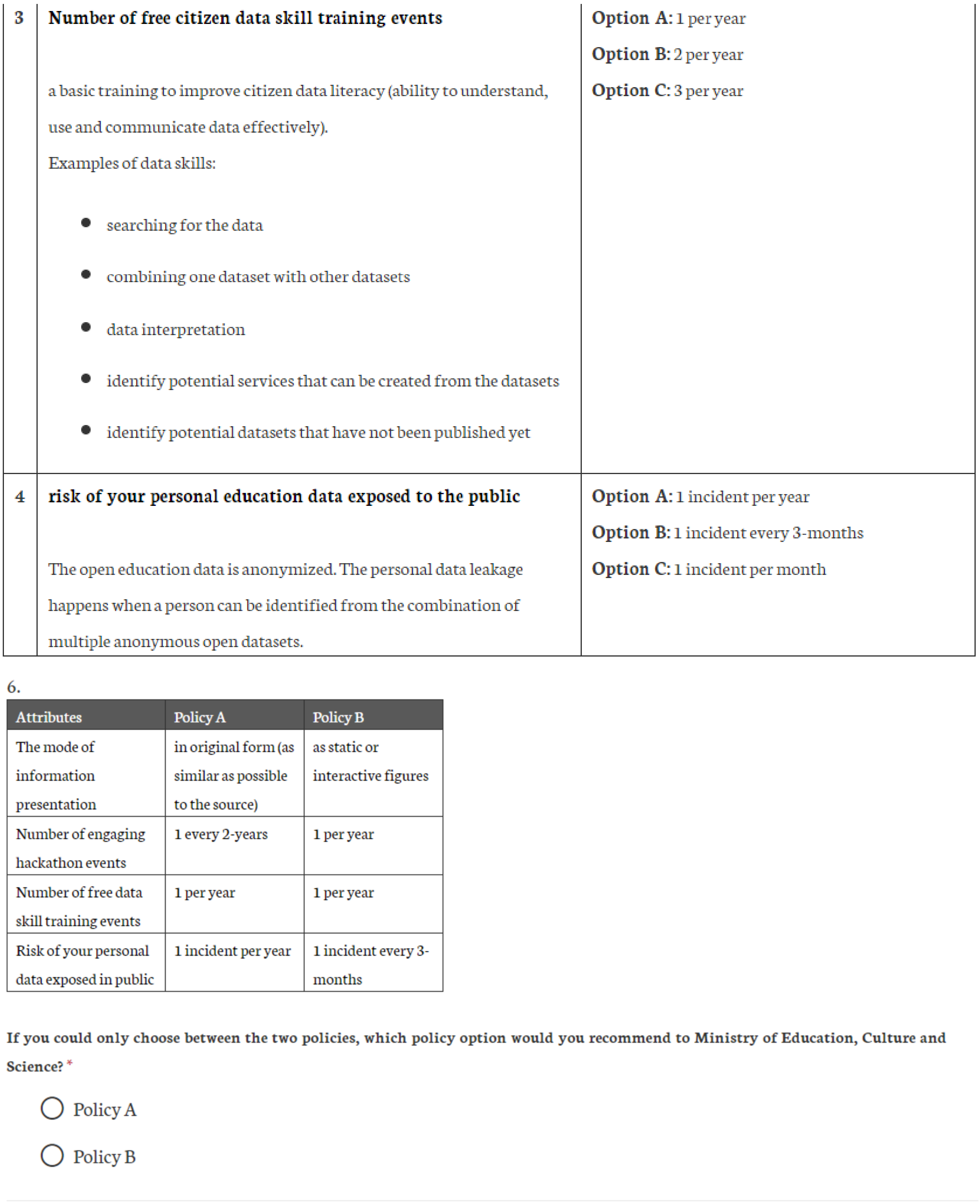
Back Next

20%

Policies for Open Education Data

In the following section, you will be presented with 9 choice situations. In each situation, you will be asked to choose between two open education data policies which are different regarding the following 4 attributes and options:

#	Attributes	Options
1	**mode of information presentation** ● Data in original form can be transformed into different forms (figures, input for other services). ● static or interactive figures communicate statistical results in the graphical forms. For example, student study satisfaction over the years. ● A service is an application created for a specific purpose. For example, studiekeuze123.nl to help students choose suitable study programs, scholenopdekaart.nl to help parents choose primary and secondary school for their children	**Option A:** in original form (as similar as possible to the source) **Option B:** as static or interactive figures **Option C:** as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)
2	**Number of free engaging hackathon events** The hackathon is organized by government to address a specific social problem using the open education data. The results can be recommendations for the government or a protype of service to address the problem. For example, Hack de Valse Start hackathon aimed to gain more insight with the help of data on how municipalities and school boards can identify and tackle inequality of opportunity in education.	**Option A:** 1 every 2-years **Option B:** 1 per year **Option C:** 2 per year



3	**Number of free citizen data skill training events** a basic training to improve citizen data literacy (ability to understand, use and communicate data effectively). Examples of data skills: ● searching for the data ● combining one dataset with other datasets ● data interpretation ● identify potential services that can be created from the datasets ● identify potential datasets that have not been published yet	**Option A:** 1 per year **Option B:** 2 per year **Option C:** 3 per year
4	**risk of your personal education data exposed to the public** The open education data is anonymized. The personal data leakage happens when a person can be identified from the combination of multiple anonymous open datasets.	**Option A:** 1 incident per year **Option B:** 1 incident every 3-months **Option C:** 1 incident per month

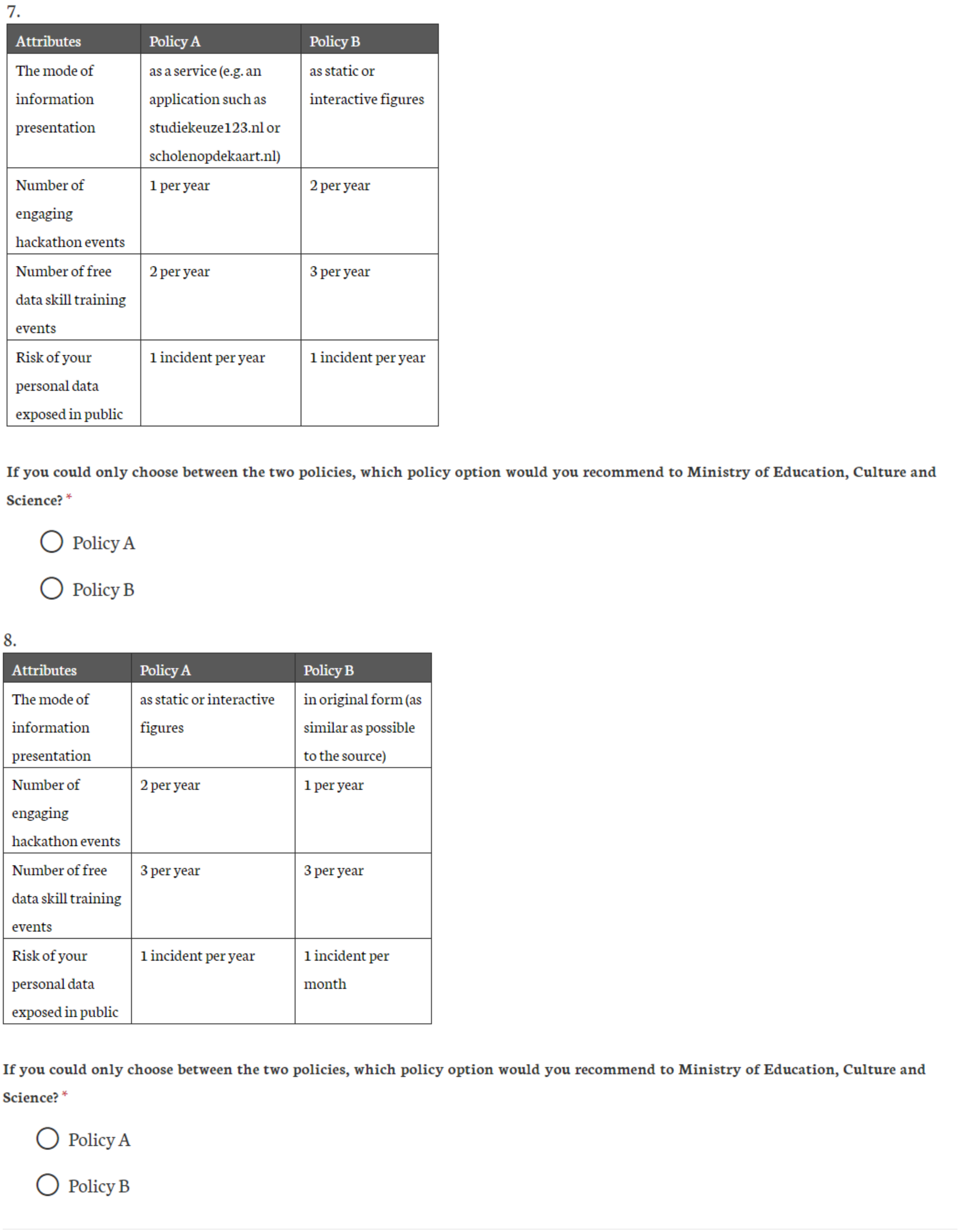
6.

Attributes	Policy A	Policy B
The mode of information presentation	in original form (as similar as possible to the source)	as static or interactive figures
Number of engaging hackathon events	1 every 2-years	1 per year
Number of free data skill training events	1 per year	1 per year
Risk of your personal data exposed in public	1 incident per year	1 incident every 3-months

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



7.

Attributes	Policy A	Policy B
The mode of information presentation	as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)	as static or interactive figures
Number of engaging hackathon events	1 per year	2 per year
Number of free data skill training events	2 per year	3 per year
Risk of your personal data exposed in public	1 incident per year	1 incident per year

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B

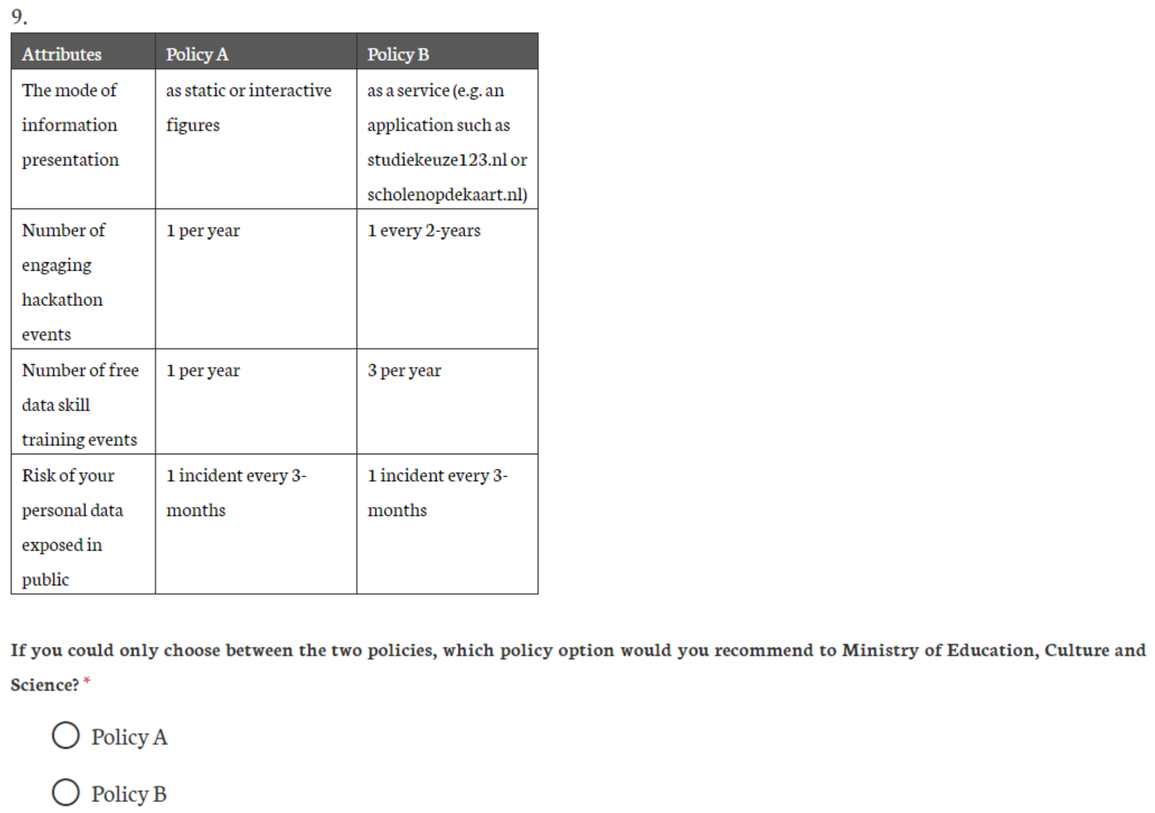
8.

Attributes	Policy A	Policy B
The mode of information presentation	as static or interactive figures	in original form (as similar as possible to the source)
Number of engaging hackathon events	2 per year	1 per year
Number of free data skill training events	3 per year	3 per year
Risk of your personal data exposed in public	1 incident per year	1 incident per month

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



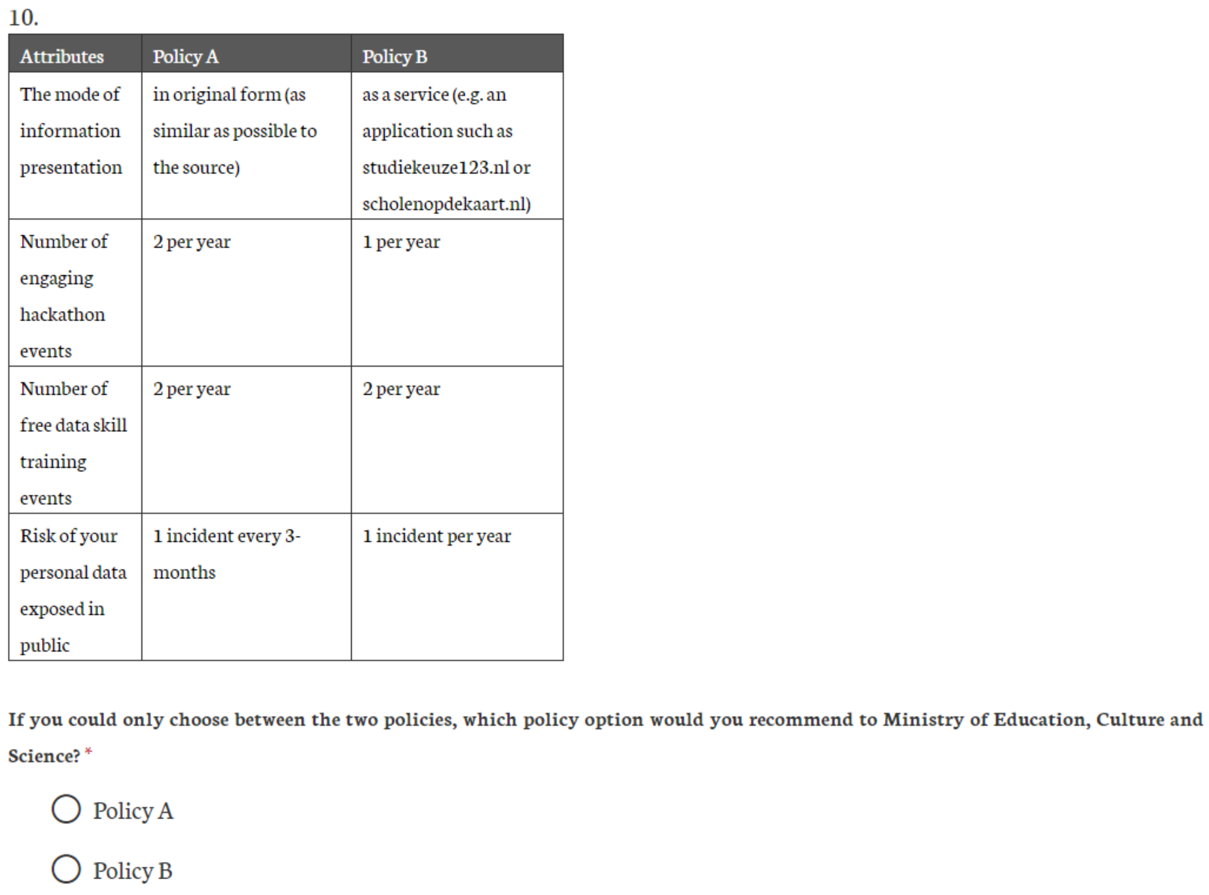
9.

Attributes	Policy A	Policy B
The mode of information presentation	as static or interactive figures	as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)
Number of engaging hackathon events	1 per year	1 every 2-years
Number of free data skill training events	1 per year	3 per year
Risk of your personal data exposed in public	1 incident every 3-months	1 incident every 3-months

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



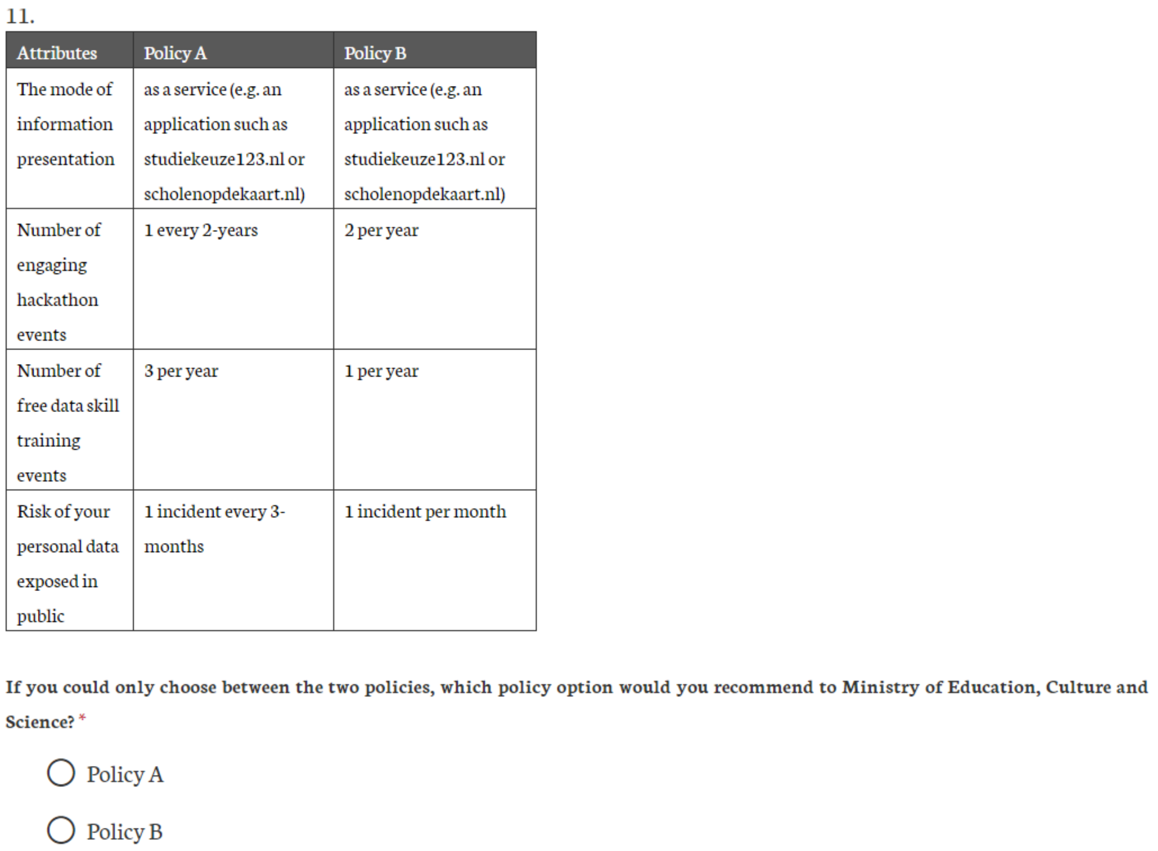
10.

Attributes	Policy A	Policy B
The mode of information presentation	in original form (as similar as possible to the source)	as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)
Number of engaging hackathon events	2 per year	1 per year
Number of free data skill training events	2 per year	2 per year
Risk of your personal data exposed in public	1 incident every 3-months	1 incident per year

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



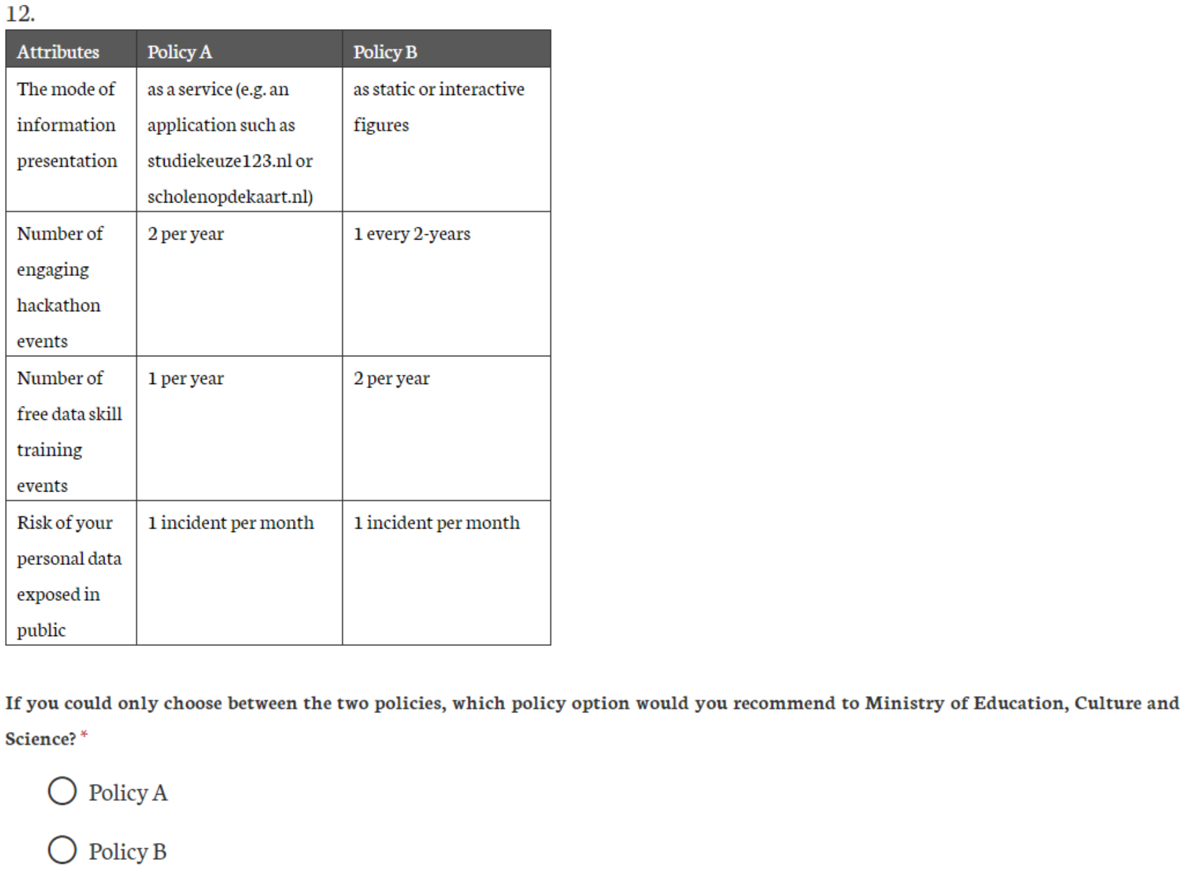
11.

Attributes	Policy A	Policy B
The mode of information presentation	as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)	as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)
Number of engaging hackathon events	1 every 2-years	2 per year
Number of free data skill training events	3 per year	1 per year
Risk of your personal data exposed in public	1 incident every 3-months	1 incident per month

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



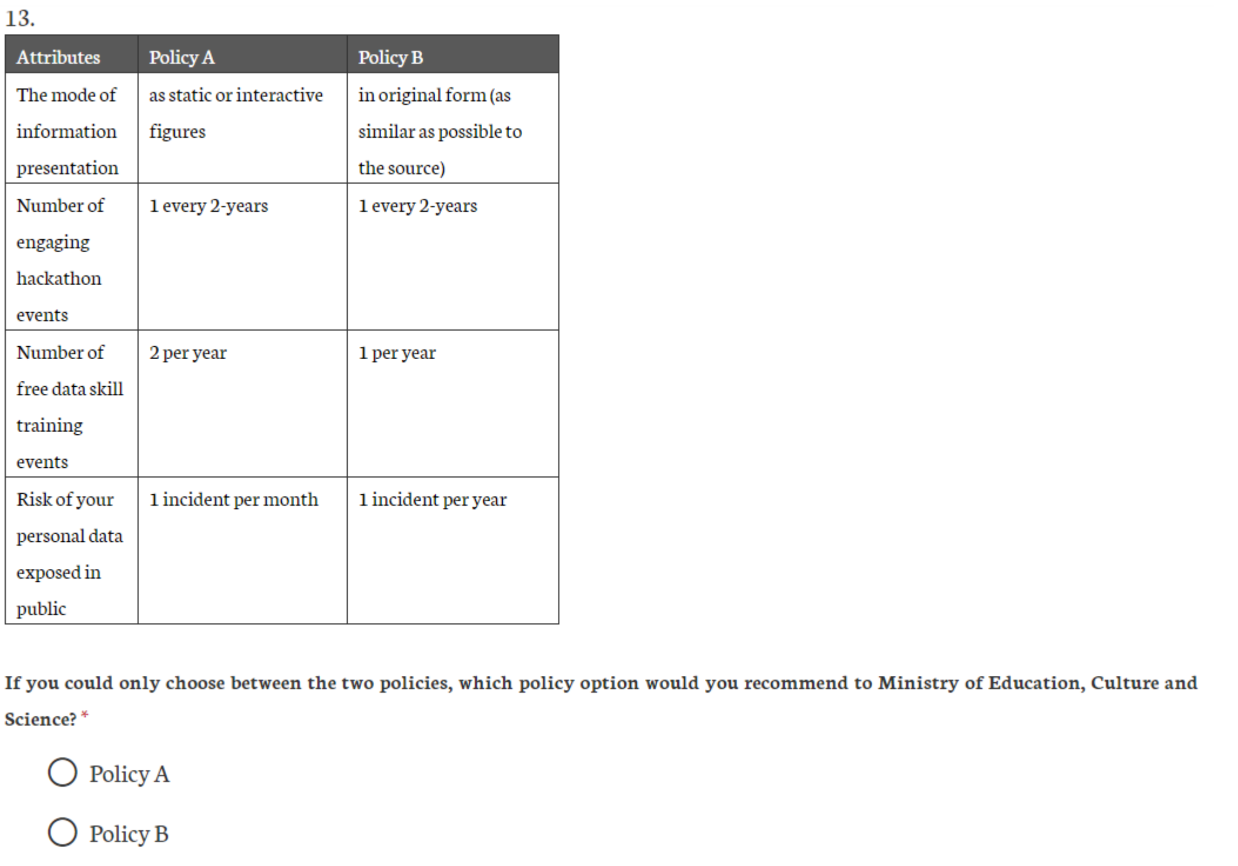
12.

Attributes	Policy A	Policy B
The mode of information presentation	as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)	as static or interactive figures
Number of engaging hackathon events	2 per year	1 every 2-years
Number of free data skill training events	1 per year	2 per year
Risk of your personal data exposed in public	1 incident per month	1 incident per month

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



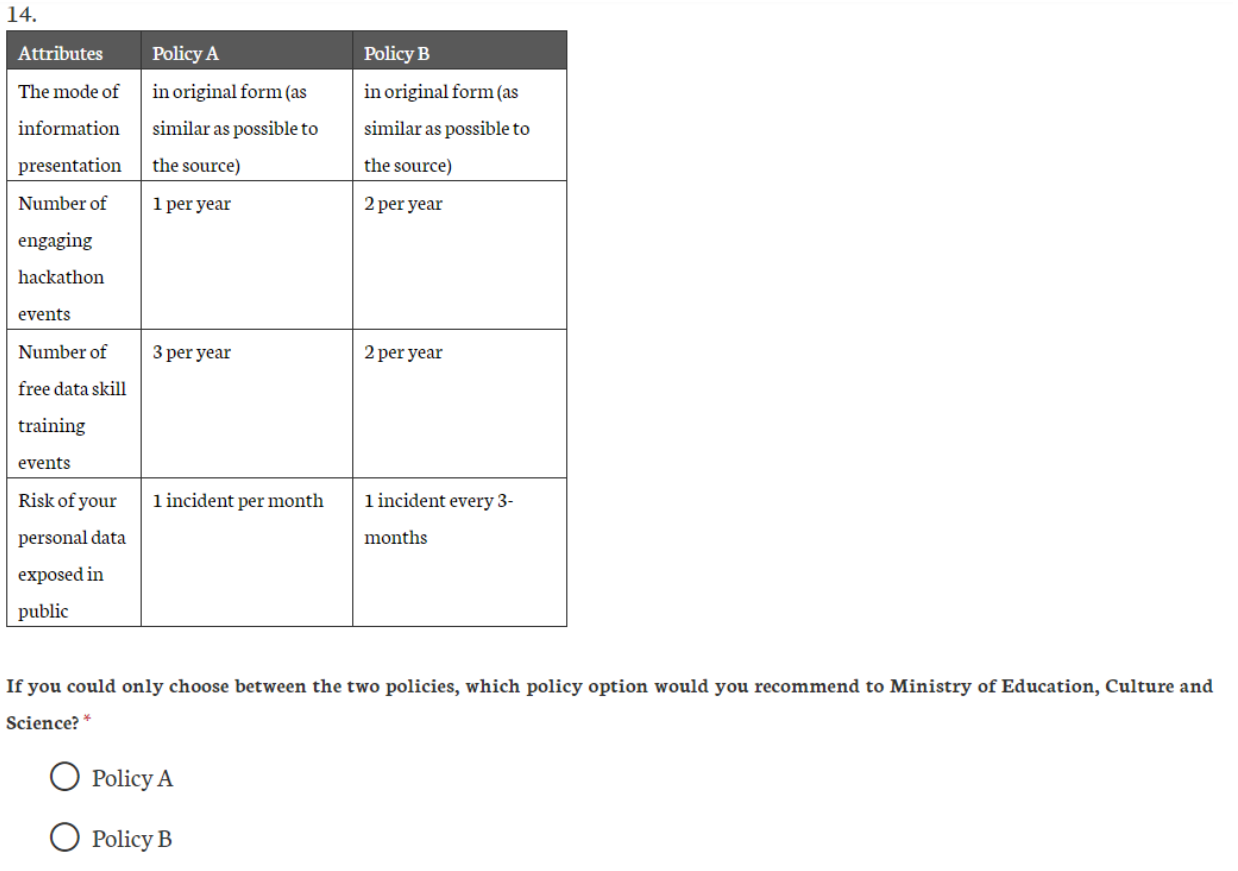
13.

Attributes	Policy A	Policy B
The mode of information presentation	as static or interactive figures	in original form (as similar as possible to the source)
Number of engaging hackathon events	1 every 2-years	1 every 2-years
Number of free data skill training events	2 per year	1 per year
Risk of your personal data exposed in public	1 incident per month	1 incident per year

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

○ Policy A

○ Policy B



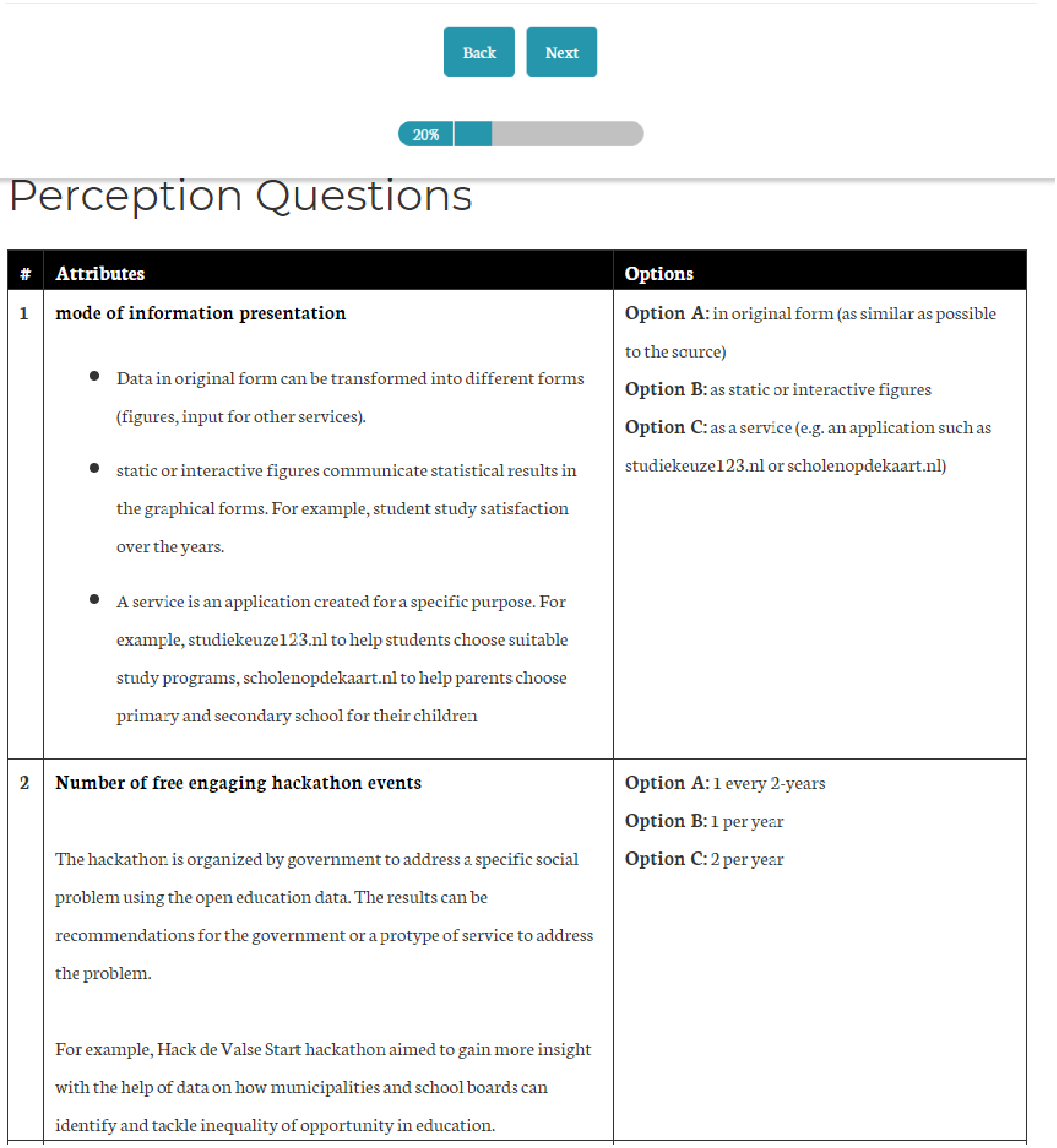
14.

Attributes	Policy A	Policy B
The mode of information presentation	in original form (as similar as possible to the source)	in original form (as similar as possible to the source)
Number of engaging hackathon events	1 per year	2 per year
Number of free data skill training events	3 per year	2 per year
Risk of your personal data exposed in public	1 incident per month	1 incident every 3-months

If you could only choose between the two policies, which policy option would you recommend to Ministry of Education, Culture and Science? *

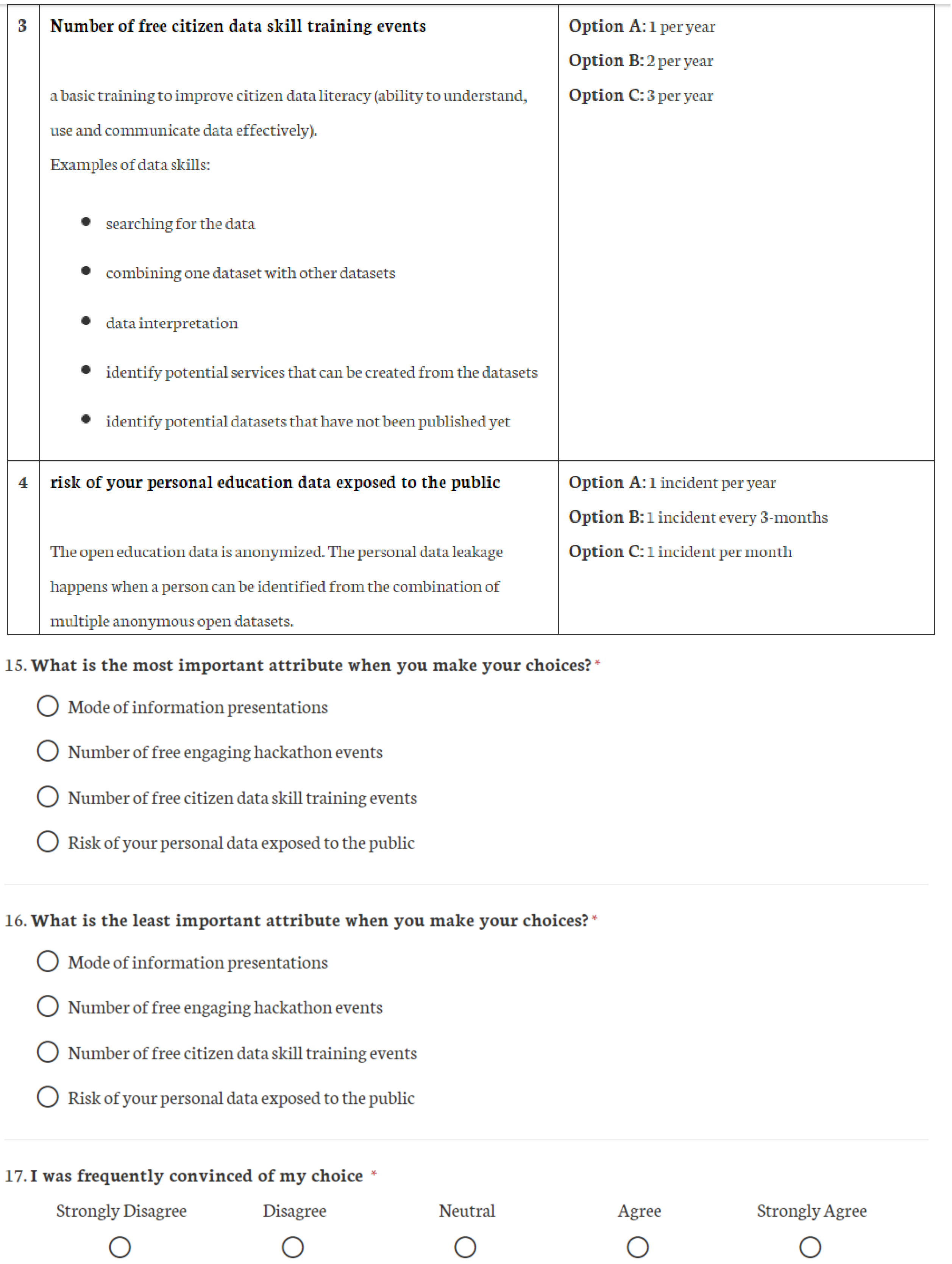
○ Policy A

○ Policy B



Perception Questions

#	Attributes	Options
1	**mode of information presentation** ● Data in original form can be transformed into different forms (figures, input for other services). ● static or interactive figures communicate statistical results in the graphical forms. For example, student study satisfaction over the years. ● A service is an application created for a specific purpose. For example, studiekeuze123.nl to help students choose suitable study programs, scholenopdekaart.nl to help parents choose primary and secondary school for their children	**Option A:** in original form (as similar as possible to the source) **Option B:** as static or interactive figures **Option C:** as a service (e.g. an application such as studiekeuze123.nl or scholenopdekaart.nl)
2	**Number of free engaging hackathon events** The hackathon is organized by government to address a specific social problem using the open education data. The results can be recommendations for the government or a protype of service to address the problem. For example, Hack de Valse Start hackathon aimed to gain more insight with the help of data on how municipalities and school boards can identify and tackle inequality of opportunity in education.	**Option A:** 1 every 2-years **Option B:** 1 per year **Option C:** 2 per year



3	**Number of free citizen data skill training events** a basic training to improve citizen data literacy (ability to understand, use and communicate data effectively). Examples of data skills: ● searching for the data ● combining one dataset with other datasets ● data interpretation ● identify potential services that can be created from the datasets ● identify potential datasets that have not been published yet	**Option A:** 1 per year **Option B:** 2 per year **Option C:** 3 per year
4	**risk of your personal education data exposed to the public** The open education data is anonymized. The personal data leakage happens when a person can be identified from the combination of multiple anonymous open datasets.	**Option A:** 1 incident per year **Option B:** 1 incident every 3-months **Option C:** 1 incident per month

15. **What is the most important attribute when you make your choices?** *

○ Mode of information presentations

○ Number of free engaging hackathon events

○ Number of free citizen data skill training events

○ Risk of your personal data exposed to the public

16. **What is the least important attribute when you make your choices?** *

○ Mode of information presentations

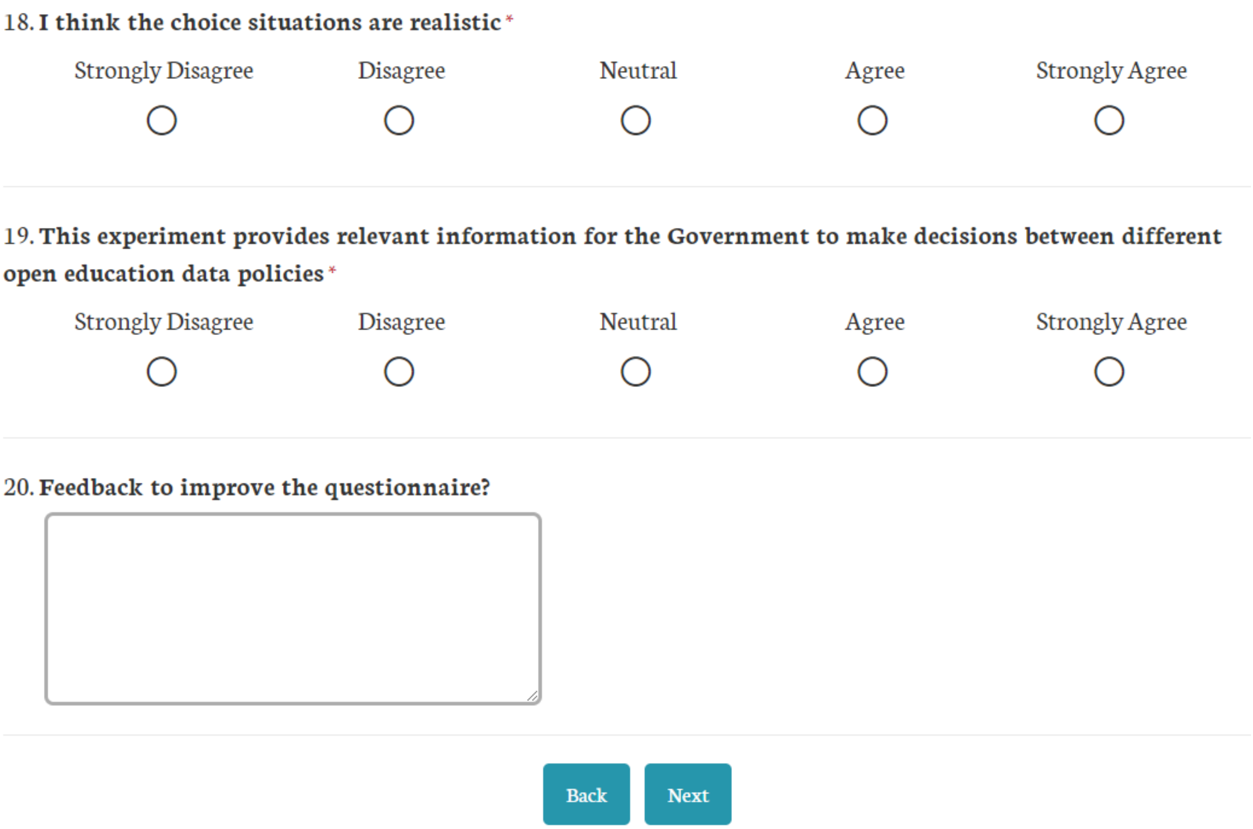
○ Number of free engaging hackathon events

○ Number of free citizen data skill training events

○ Risk of your personal data exposed to the public

17. **I was frequently convinced of my choice** *

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○



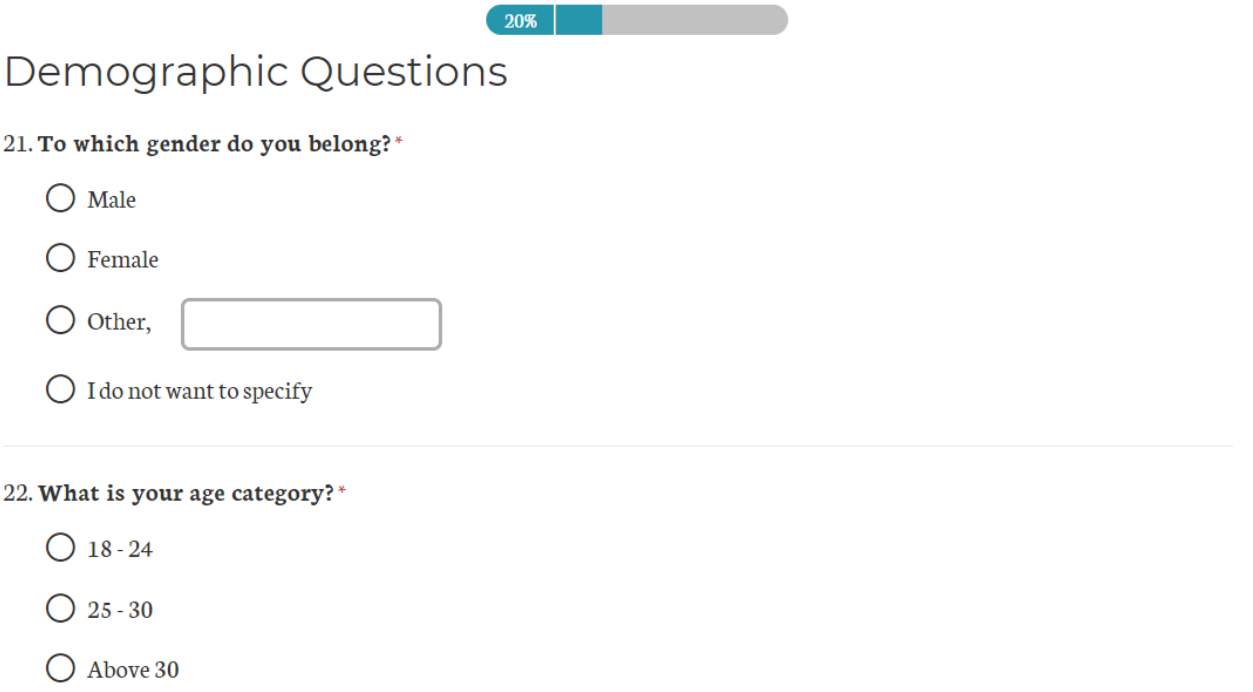
18. **I think the choice situations are realistic** *

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

19. **This experiment provides relevant information for the Government to make decisions between different open education data policies** *

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

20. **Feedback to improve the questionnaire?**



Demographic Questions

21. **To which gender do you belong?** *

○ Male

○ Female

○ Other,

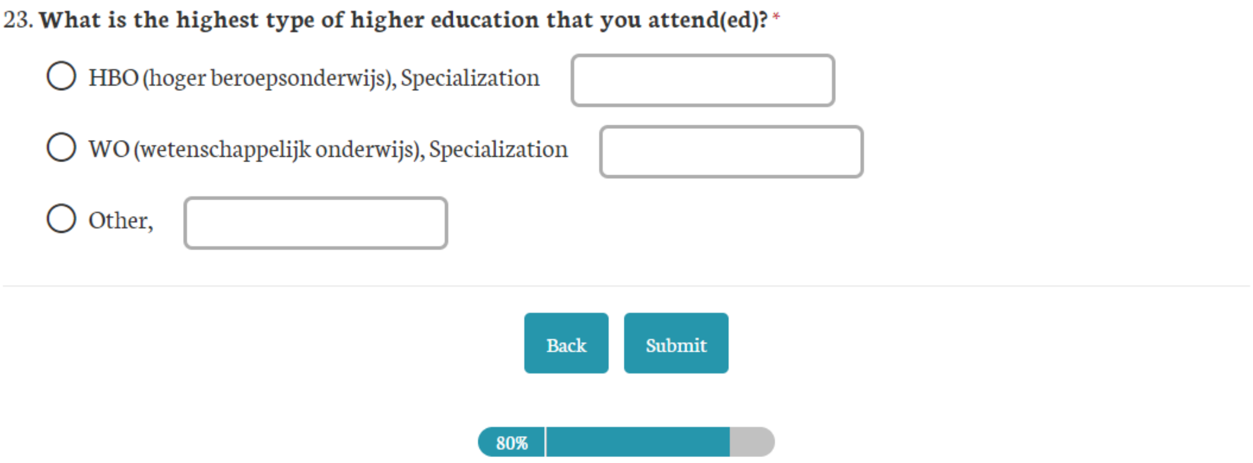
○ I do not want to specify

22. **What is your age category?** *

○ 18 - 24

○ 25 - 30

○ Above 30

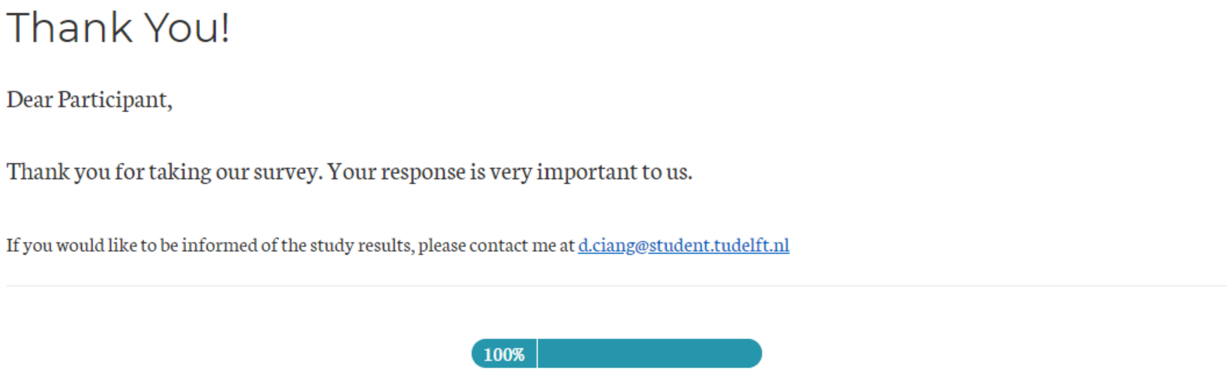


23. **What is the highest type of higher education that you attend(ed)?** *

○ HBO (hoger beroepsonderwijs), Specialization

○ WO (wetenschappelijk onderwijs), Specialization

○ Other,



Thank You!

Dear Participant,

Thank you for taking our survey. Your response is very important to us.

If you would like to be informed of the study results, please contact me at d.ciang@student.tudelft.nl