UNIVERSITY OF CAMBRIDGE

Bennett Institute for Public Policy Cambridge

WORKING PAPER

Potential social value from data: an application of discrete choice analysis

By

Diane Coyle[a*]

Annabel Manley[a*]

[a]Bennett Institute for Public Policy, University of Cambridge, UK

5 August 2021

*We are grateful for funding for this work from the Omidyar Network. We'd like to thank Richard Heys, Angela Potter and Susan Williams at the Office for National Statistics for their very helpful insights; Marco Felici, Sam Gilbert, and Rehema Msulwa for their comments on an earlier draft. We are also very grateful to the AEA members who took part in the survey. We are of course responsible for any errors and omissions.

Abstract

Data is understood to be a key digital economy resource and yet one that is difficult to value, as there are relatively few market prices for data. In addition its non-rival nature and pervasive externalities mean there will also be a wedge between any market price and the social value of the data. This paper proposes an approach to estimating the social value of certain kinds of data sets using a discrete choice approach, and a methodology that could for example be used by public bodies to understand the value of their data investments. We test this approach on a specific data set as proof of concept.

JEL Codes: C80, C83, D60

Key words: data, discrete choice, digital

Introduction

There is great interest in understanding the value of data, and different empirical approaches have begun to be explored, in the academic literature and in commercial contexts. However, the existence of informational externalities and the non-rival character of data immediately imply that private markets and uses alone will not deliver an economically efficient social availability of data, nor will any market prices (if they exist) reflect the social value. What's more, the value of any given data set is also fundamentally determined by the value of the uses to which it can be put, which are likely unknown until after the fact.

In this paper we apply a form of discrete choice experiment, applying it to the question of how public sector controllers of data could estimate the potential social value other users would gain from the data and also help identify which attributes of the data have particular value to users. This could inform decisions such as whether public bodies should sell data access to the commercial sector to cover their costs, or how much to invest in maintaining open access data bases by estimating a social benefit for a cost-benefit analysis. We use a commercially available discrete choice analysis software package, conjoint.ly, widely used in marketing and show how it can elicit preferences and marginal values capturing, at least to some degree, the data externalities about which there are, at present, few empirical insights. The method is straightforward to use and could be readily applied in practical contexts.

The question we want to address is: what is the potential social value of data that could arise from its use by individuals, innovators or civic organisations? Data policy decisions will depend on policymakers having estimates of the scale of such potential benefits from access to data, given the existence of trade-offs. For example, they may need to compare the potential social benefit of wider free access to geospatial or transportation data – perhaps through the innovations this could enable or the time it would allow people to save – with the costs of collecting and maintaining the data, and the potential loss of commercial income from sale of data to commercial entities, and thus restricted access. At present, there is no standard empirical approach to the measurement of this potential external value. Indeed, understanding the full impact of data on social welfare is described by Pei (2021) as "the grand challenge," given the trade-offs involved.

This paper takes a small step toward tackling the grand challenge. The discrete choice analysis produces an estimate of the economic value of data in terms of its potential contribution to social welfare by eliciting willingness to pay, a measure of consumer surplus. Discrete choice analysis is one form of stated preference method, widely used in marketing research, to uncover people's preferences for different product characteristics and enable pricing decisions for new goods that previously were not on the market. The software we use, conjoint.ly, takes an approach which is consistent with standard consumer demand theory. The method is based on standard utility theory in economics, with the values individuals assign to goods in surveys that present them with 'bundles' of product attributes being revealed by the choices they state.

There is a large literature on contingent valuation methods, which is widely used to understand consumer valuations and preferences in contexts where there are no monetary prices, such as environmental or cultural goods (see e.g. Carson, Flores & Meade, 2001 and McFadden & Train, 2017 for surveys). While economists always prefer to calculate values based on observed market prices and quantities where possible, some (e.g. Blinder 1991) defend the need to use interview or survey techniques in contexts where economics is otherwise unable to provide any preferred method for empirical estimation – as is the case with many non-monetary, non-rival goods, from a clean environment to data. While data is a 'club good' rather than a pure public good, as access can be restricted through technical and legal means, estimating its full social value cannot rely on prices set in data markets. We discuss the limitations of our approach below.

At present, much data is held by companies or organisations for their own commercial use, and there is a rapidly-growing literature (as well as commercial practice) on how they might realise more value from better using their data. For example, there is now important work exploring the use of a 'data Shapley' approach to estimate the value of specific data sets to their owners or controllers in terms of the potential for improved outcomes in their objective function, such as profit (e.g. Arrieta Ibarra et al 2020, Ghorbani & Zou 2019, Jia et al 2020). There is also work on designing private markets for data (e.g. Galperti et al 2021, Koutrompis et al 2020). A number of authors suggest methods based variously on stockmarket valuations, market prices, revenues or other business metrics, or costs (e.g. Ker & Mazzini 2020, Birch et al 2021, Coyle & Li 2021).

Yet for public controllers of data concerned to maximise social welfare, methods based on realised financial values in market transactions are insufficient. For additional social value could potentially be gained from more data collection, wider access, or the scope to join information from different data sets with varying types of data records (noting also the need to manage the negative externalities of potential privacy loss and security breaches) (Coyle & Diepeveen 2021).

We take total social welfare as the objective function, and the discrete choice experiment we conduct is in the tradition of the growing literature on using contingent valuation methods to estimate the value of 'free' (zero monetary price) digital goods (e.g. Allcott et al 2020, Brynjolfsson et al 2020, Coyle & Nguyen 2021).

As far as we are aware our specific approach in this paper has been used to date in this context of data valuation only in a pilot study by the Office for National Statistics, seeking to understand the value of official statistics (Williams 2021). A similar method has been used to devise a means of market definition in zero price digital markets by estimating stated preference price elasticities (Nakamura & Ida 2021). The United Nations Economic Commission for Europe (UNECE 2018) has called on national statistical agencies to measure the value of official statistics in enabling better decision-making by governments, businesses and households. Since many official statistics datasets are accessible under public license with no monetary price, this method could potentially be used to estimate how much individuals would be willing to pay if they were no longer able to access the data. It therefore can enable public bodies to understand better the societal value of the data they control, and hence to make better-informed decisions themselves about access, charging and data investment policies.

Methodology

To conduct our analysis, we used the software conjoint.ly, one of a range of commercially-available software packages which automates the experimental design, including which combinations of data attributes to present to survey respondents. The analysis is a discrete choice experiment using a hierarchical Bayesian method to estimate a multinomial logit model of the decision parameters (Rossi & Allenby 2003). The method pulls repeated random samples from a distribution of utilities to infer values. The benefit of this approach is that it accounts for all the heterogeneity across individuals in the market and uses all available information, not static priors. It is more reliable the larger the sample; conjoint.ly suggests a minimum of 100 respondents, which seems low by normal statistical inference standards although the usual context for its use is rapid marketing decisions.[1] The software uses 'choice-based conjoint', or in other words discrete choice experimentation, which gives survey respondents multiple profiles

[1] https://conjointly.com/guides/conjoint-technical-notes/

and asks them to choose their preferred combination, compared to traditional conjoint analysis in marketing where each individual option is simply ranked (Louviere, Flynn & Carson 2010).[2]

The premise of the approach is that a product is defined by its 'attributes,' or in other words it can be broken down to component features. The different values of an attribute are then defined as 'levels.' For example, applying this to a physical product, one potential attribute would be its colour, and the levels could be 'black', 'white', 'red', etc. The survey then generates hypothetical products made up of combinations of these levels from each attribute along with prices to determine which levels are most attractive, the relative importance of each attribute, and the price the survey respondents are willing to pay for various features. Where one of the attributes tested is price, the method can be used to generate willingness to pay estimates across the distribution of heterogeneous preferences of respondents. The calculations therefore indicate social welfare but with limitations, notably assuming linear price-preference schedules between price levels.

Survey

For this initial analysis, we invited participation by approximately 8,000 members of the American Economic Association (AEA) who had previously agreed to be sent requests to take part in surveys. This service by the AEA is new and reflects the growing interest survey methods in economics (for example, Harvard's Social Economics Lab). Our sample was therefore not selected purposively nor is it representative. We received 401 responses, above the 100 minimum number of responses recommended by conjoint.ly for a survey with our selected number of attributes and levels, but smaller than desirable for robustness.

We opted for this approach, rather than for instance a representative sample, because the method requires people to have a reasonable understanding of the products being presented, including how variation in the levels of an attribute affect how a dataset can be used. While the attributes of many physical products are easy to understand, such as colour or size, the attributes of datasets are more technical, so that there is a higher risk of cognitive overload from a given number of

[2] https://towardsdatascience.com/choice-based-conjoint-analysis-dafcff135c2. In the marketing literature the general term 'conjoint analysis' is often used to refer to both discrete choice experimentation or 'choice-based conjoint' and more restrictive traditional conjoint or 'partworth' analysis.

attributes. The AEA's members are more likely than the general population to be familiar with the concepts and attributes used, and are therefore more likely to be able to accurately represent their own preferences for different combinations of attribute levels.

We selected a dataset likely to be familiar to the participants, the World Bank's World Development Indicators (WDI). In our sample, 231 people said they had used the data set. This is open access, so to understand the value its users place on it, we presented the hypothetical situation that people could be charged for access. The WDI dataset is a collection of development indicators compiled from officially recognized international sources, with around 1,400 time series indicators for over 250 economies and country groups, going back 60 years.[3] The WDI was chosen for the experiment due to the likelihood that the sample would be familiar either with the dataset directly, or certainly the standard variables it contains, such as GDP, inflation, unemployment, etc. The comprehensive nature of the dataset and its wide use in economics, also ensure its open access is likely to have non-negligible value, which is beneficial for testing the method. However, it is worth noting that datasets with familiar aggregate statistics are likely to have different drivers of value than microeconomic (or even 'big') datasets, whose users will want detailed individual-level information for a range of different uses. Our approach is more suited to the former type of data.

Respondents were first asked about their country of residence and stage of career, which may affect how reliant they are on open access datasets (as early career researchers are less likely to have funding to purchase data). They were then asked questions to elicit their attitude toward open data and what they consider its most important attributes. They were also asked if they had used the WDI previously.

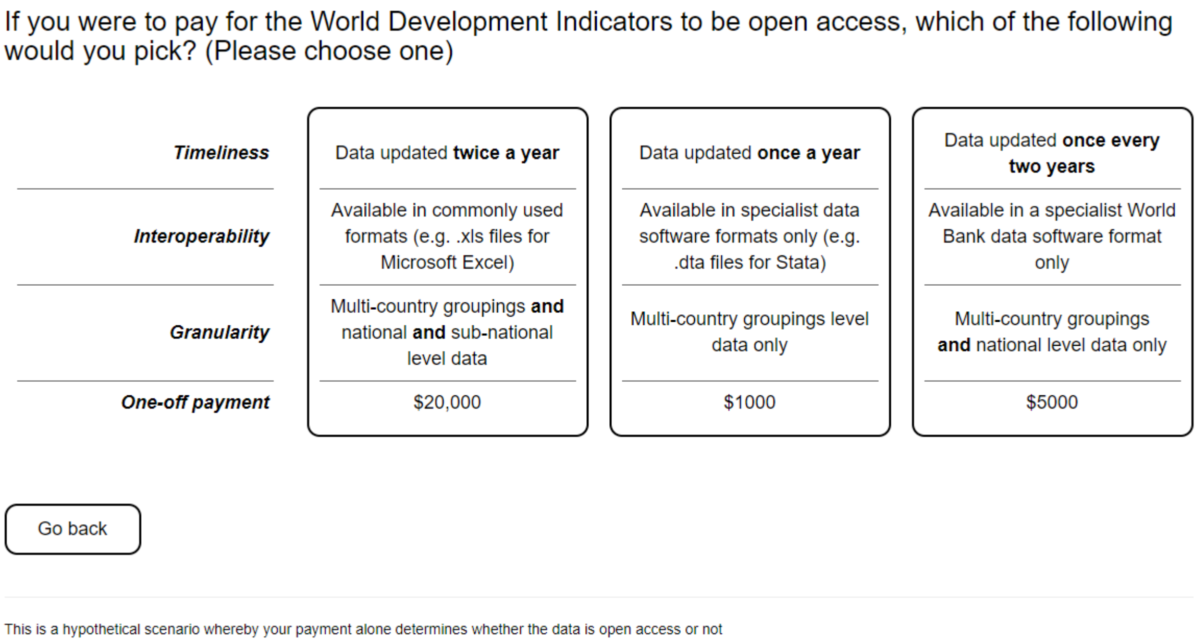
Before the survey was launched for the participants, we also presented an open text response question asking for the three attributes of a dataset they considered would most determine its economic value. This was done before they saw our pre-selected attributes or dataset and so before they were affected by our survey design choices.

They were then presented with the hypothetical situation of being charged for access to the data and were asked a series of eight questions generated by the software, an example of which is shown in Figure 1.

[3] https://datatopics.worldbank.org/world-development-indicators/

We restricted the number of attributes to three, plus a hypothetical price, in order to limit the difficulty of comparison for the respondents; more can become confusing. The literature highlights many characteristics that can determine the value of data, and there are various taxonomies of data (Coyle et al 2019), and so a limited choice will inevitably omit features that might be considered important. We selected three attributes that a priori to us as economists seemed relatively important, and are easy to understand and identify in the context of the WDI. These are the timeliness (how frequently the data is updated), interoperability (how easy it is to download and use), and granularity (how detailed the data gets in terms of its coverage). These were each split into three 'levels'. We tried to ensure that the levels match to commonly found or used reasonable anchors, e.g. price points were set with reference to plausible research budgets.

Figure 1: Example survey format



If you were to pay for the World Development Indicators to be open access, which of the following would you pick? (Please choose one)

Timeliness	Data updated **twice a year**	Data updated **once a year**	Data updated **once every two years**
Interoperability	Available in commonly used formats (e.g. .xls files for Microsoft Excel)	Available in specialist data software formats only (e.g. .dta files for Stata)	Available in a specialist World Bank data software format only
Granularity	Multi-country groupings **and** national **and** sub-national level data	Multi-country groupings level data only	Multi-country groupings **and** national level data only
One-off payment	$20,000	$1000	$5000

This is a hypothetical scenario whereby your payment alone determines whether the data is open access or not

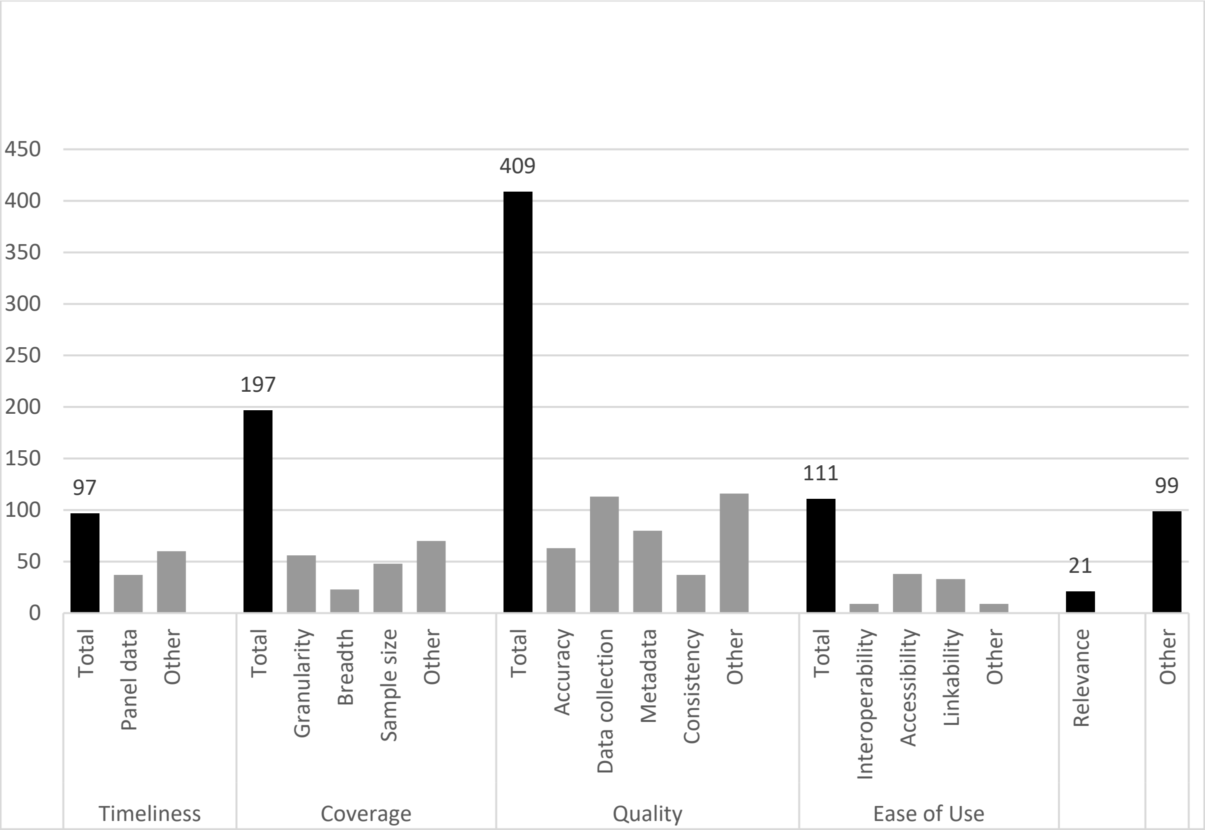
Results

We begin by reporting responses to the open question about which three data attributes respondents considered most important. Not all of the 401 participants gave exactly three characteristics, and the total number of codified characteristic responses was 934. The responses fit into five broad themes.

- First, responses related to timeliness or data frequency, including whether it was a panel/ longitudinal dataset.
- Second, responses related to the size and coverage of the data, including its granularity, breadth, and sample size.
- Third, responses related to the quality of the dataset, including accuracy, data collection methodology, the quality of the accompanying documentation or metadata, and the consistency of the data collection.
- Fourth, responses related to the ease of use of the data, including its interoperability, accessibility or availability, and linkability to other datasets.
- Fifth, responses related to the relevance of the data to their work.

Finally, 10.5% of the responses did not fit into these categories, including comments about content of a dataset or what sort of variables should be included. The frequency of the responses is shown in Figure 2, with the black bars showing the total responses for each of the broad themes, and the grey bars showing their decomposition within them.

Figure 2: Frequency of characteristics mentioned by participants.



The results broadly supported our selection of attributes for the discrete choice analysis. Our prior selection omitted the range of different 'quality' variables, however, as well as the less-frequently cited relevance attribute. The results also produce a goodness of fit measure.[4] Both for the entire sample and for the previously mentioned specified subsamples, this is around 75%, indicating a strong fit between the attributes and the variation shown in the survey results. This confirms that the choices made by participants reflect their preferences rather than simply arbitrary responses.

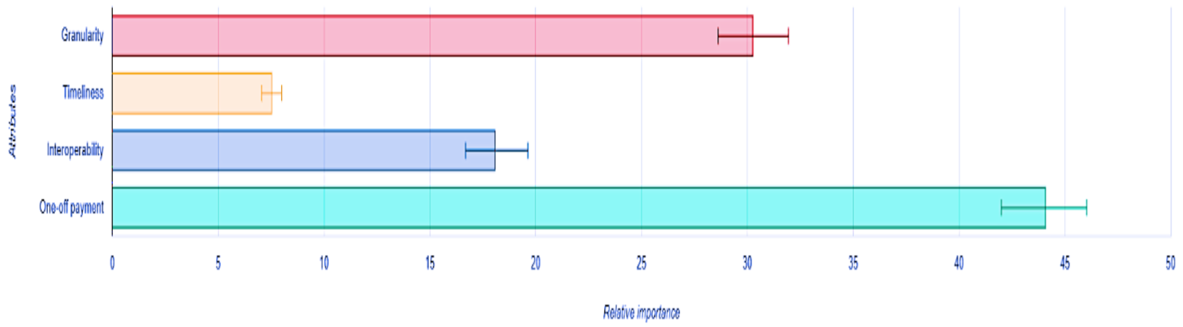
The software outputs the relative importance of each attribute surveyed, the relative value of each level of each attribute, the marginal willingness to pay, and the goodness of fit. It also allows for segmentation of the sample using the initial questions we asked our survey respondents.

Figure 3 illustrates the relative importance to the full sample of respondents of each attribute (with 95% confidence intervals). The price of the data access emerged as the most important attribute, explaining the largest proportion of decisions, followed by the granularity of the data, then interoperability, and finally timeliness. However, an important caveat is that relative importance of the attributes does depend on the range in levels for each attribute chosen in the survey design. For example, under timeliness the choices that participants were presented with were twice a year, once a year, and once every two years. Participants may have felt all three options were sufficiently frequent for their needs and so placed little importance on timeliness in their decisions. However, if the levels presented were altered to be, say, one-year, five-year, and ten-year frequency of data, the relative importance of this attribute might increase.

A question for future users of this method will be how to pre-test attribute selection and levels, including whether there are threshold levels for certain attributes beyond which their importance for users would affect the relative importance results. For example, in this case either high or low frequency timescales might prompt different results.

[4] McFadden's pseudo-R^2

Figure 3: Relative importance by attribute

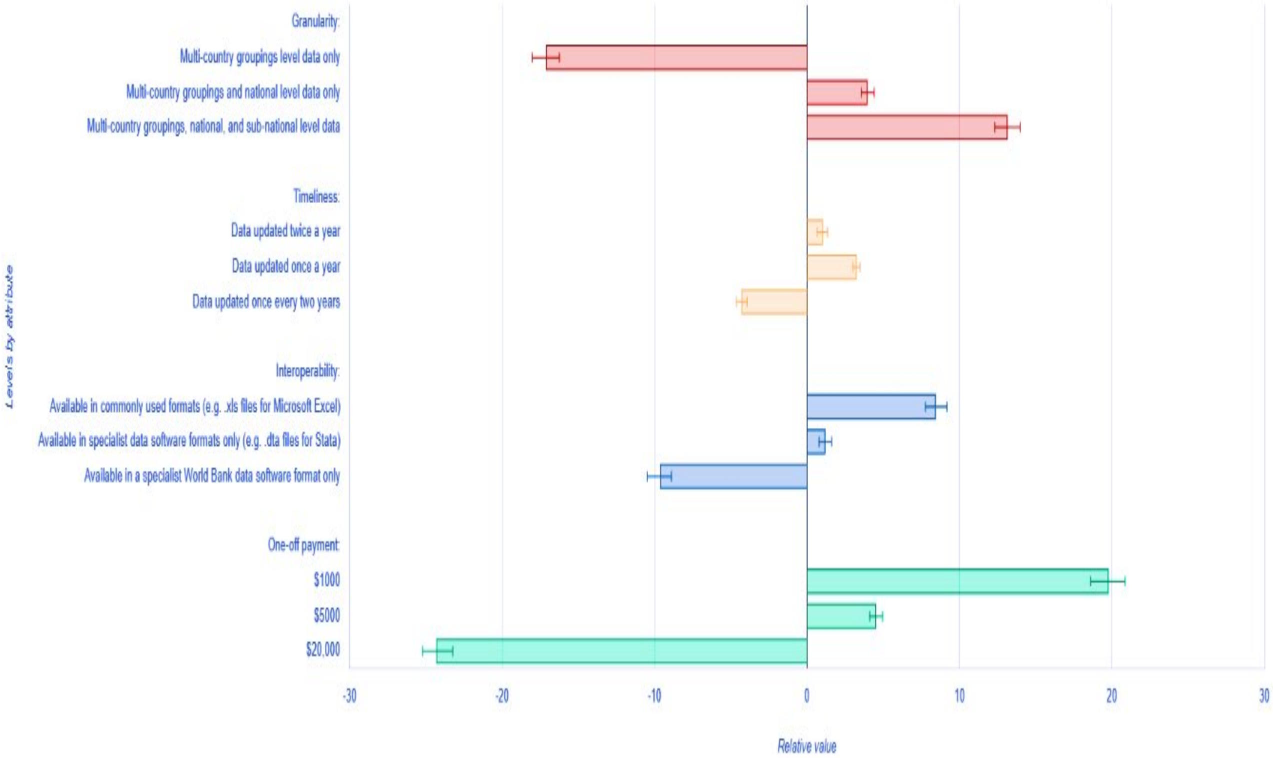


Granularity: 30%; Timeliness: 8%; Interoperability: 18%; One off payment: 44%

The relative value by level of each attribute is shown by Figure 4. The chart is scaled so that the sum of the relative values of the levels for each attribute is equal to zero, and therefore the size of the bars is not comparable across attributes. These results indicate which levels within an attribute are most preferred. Three of the attributes followed patterns as expected, with higher degrees of granularity, higher levels of interoperability, and lower prices preferred. Timeliness preferences, however, were nonlinear, with the highest preference for annual data rather than for data every six months. There are various possible reasons for this, such as other related datasets conventionally being published on an annual basis, or high serial correlation in the type of data included.

It is also notable that even though the sample surveyed consisted of economists likely to have both access to and the skills to use specialist software, data being available in commonly used simple formats, such as Microsoft Excel, was significantly preferred to data only available in more specialised software.

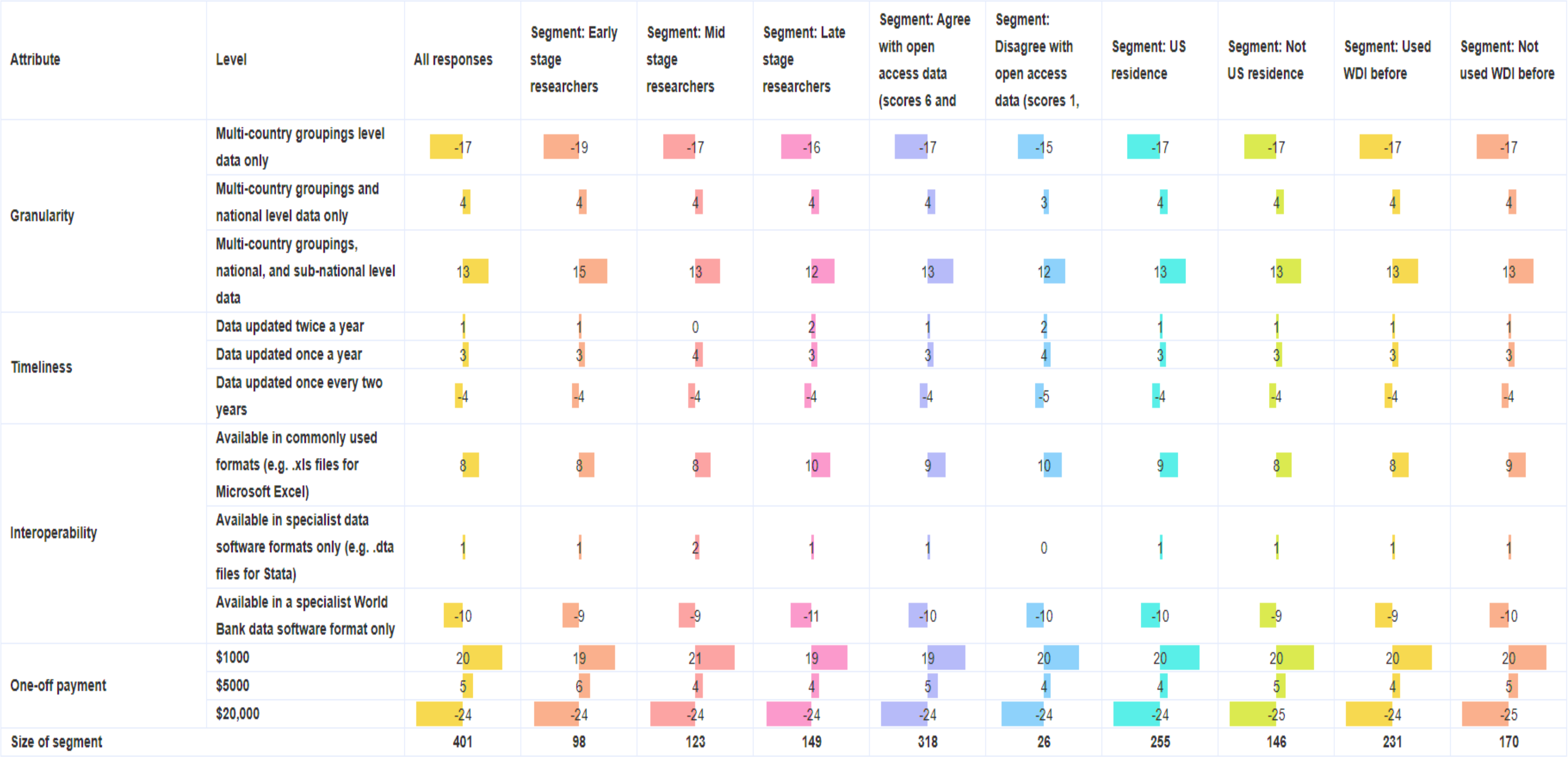
Figure 4: Relative value by level



As noted, we had included several questions prior to the discrete choice survey to allow segmentation of the sample. This included by country of residence, stage of career (early, mid, or late stage researcher), and a 7-point scale question asking participants to what extent they agreed with the statement that “All officially-produced, anonymized datasets should be open access.” We also asked the respondents whether they had used the World Development Indicators before, as this may have affected how much they valued it. Unsurprisingly given the sample, the majority (64%) resided in the US, and so to preserve reasonable sample sizes, we divided the respondents into those residing in the US or not. Similarly, a large majority (79%) chose the top two options on the 7-point scale, and so we divided the sample by those who strongly or quite strongly agreed with the statement (score of 6 or 7), and those that did not (scores 1, 2, and 3) (6%). These four ways in which to divide the sample generated results for nine subsamples.

Table 1 shows the relative value by levels for each of the nine subsamples, with the ‘all responses’ column showing the same information as is in Figure 4. Surprisingly, there is no large difference in relative values by level across any of these subsamples (including between users and non-users of WDI, by country, or by career stage), shown by reading across the columns in Table 1. We explored the individual results in detail to check this finding (see Appendix), and there is indeed relatively little variation among individual coefficients, likely due to this being a database with variables well-understood by economists, but this would not always be the case.

Table 1: Relative value by level of subsamples



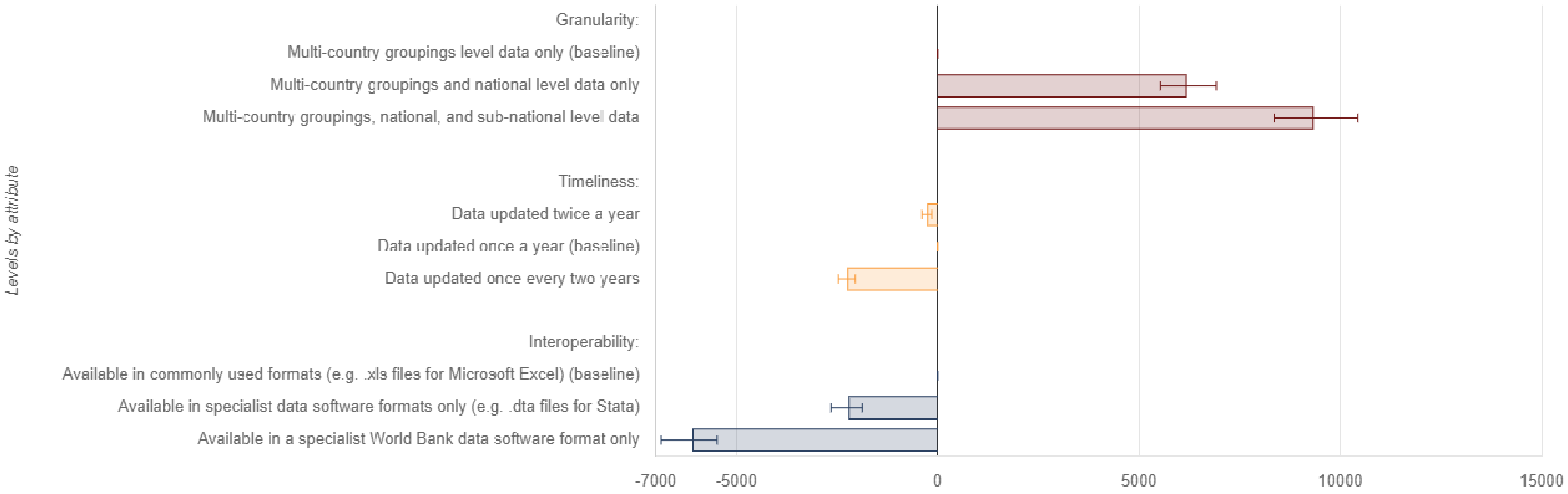
Attribute	Level	All responses	Segment: Early stage researchers	Segment: Mid stage researchers	Segment: Late stage researchers	Segment: Agree with open access data (scores 6 and	Segment: Disagree with open access data (scores 1,	Segment: US residence	Segment: Not US residence	Segment: Used WDI before	Segment: Not used WDI before
Granularity	Multi-country groupings level data only	-17	-19	-17	-16	-17	-15	-17	-17	-17	-17
	Multi-country groupings and national level data only	4	4	4	4	4	3	4	4	4	4
	Multi-country groupings, national, and sub-national level data	13	15	13	12	13	12	13	13	13	13
Timeliness	Data updated twice a year	1	1	0	2	1	2	1	1	1	1
	Data updated once a year	3	3	4	3	3	4	3	3	3	3
	Data updated once every two years	-4	-4	-4	-4	-4	-5	-4	-4	-4	-4
Interoperability	Available in commonly used formats (e.g. .xls files for Microsoft Excel)	8	8	8	10	9	10	9	8	8	9
	Available in specialist data software formats only (e.g. .dta files for Stata)	1	1	2	1	1	0	1	1	1	1
	Available in a specialist World Bank data software format only	-10	-9	-9	-11	-10	-10	-10	-9	-9	-10
One-off payment	$1000	20	19	21	19	19	20	20	20	20	20
	$5000	5	6	4	4	5	4	4	5	4	5
	$20,000	-24	-24	-24	-24	-24	-24	-24	-25	-24	-25
Size of segment		401	98	123	149	318	26	255	146	231	170

These initial results concerning the relative importance of attributes and relative value by level are both a type of result called 'partworth' utilities, with the former referred to as the "attribute partworth" and the latter the "level partworth." These are the marginal rates of substitution between attributes or levels, averaged across the sample. As in standard consumer preference theory, these partworths can be monotonically rescaled without changing the results. Thus the absolute value of the partworths has no interpretation, and the outcomes represent ordinal results. For example, data with both national and sub-national level statistics is preferred to having just the subnational data alone, but you cannot say that it is liked over three times as much simply because the level partworths are 13.18% and 3.6%. The numbers reflect only that that the sum of all level partworths for a given attribute are set equal to zero.

Our question concerns the potential of discrete choice analysis for understanding the social value of access to data in monetary terms. For this, we need the marginal rate of substitution between the levels of an attribute and the price levels, that is the ratio of two partworths where one attribute is price. These results for the marginal willingness to pay for each level of an attribute are shown Figure 5. This may be important for decisions such as the costs and benefits of making specific datasets, or parts of them, open, or for setting prices for access. For example, in the kind of policy trade-off discussed earlier, the results of the discrete choice analysis could point to making certain aggregations over time or categories open while charging for access to more detail, informed by the preferences of users of different kinds.

The results of Figure 5 are relative to selected baselines: of grouped country data only for the granularity attribute; data updated once a year for the timeliness attribute; and data available in common formats for the interoperability attribute. The "Generic" template we chose from the software does not include interaction variables and therefore the marginal willingness to pay results within an attribute are not sensitive to the baselines of the other attributes. These results reflect both the relative importance by attribute and relative value by level results, demonstrating the relative rankings among the latter but the absolute magnitudes of the former. Crucially, the magnitudes of these results therefore have meaning and are comparable. Assuming no framing effects, they should also hold as the survey design changes.

Figure 5: Willingness to pay, dollars



Granularity: National level data ($6.19k), subnational level data ($9.34k); Timeliness: Twice a year (-$265.45), once every two years (-$2.25k); Interoperability: Specialist data software (-$2.21k), specialist World Bank software (-$6.09k)

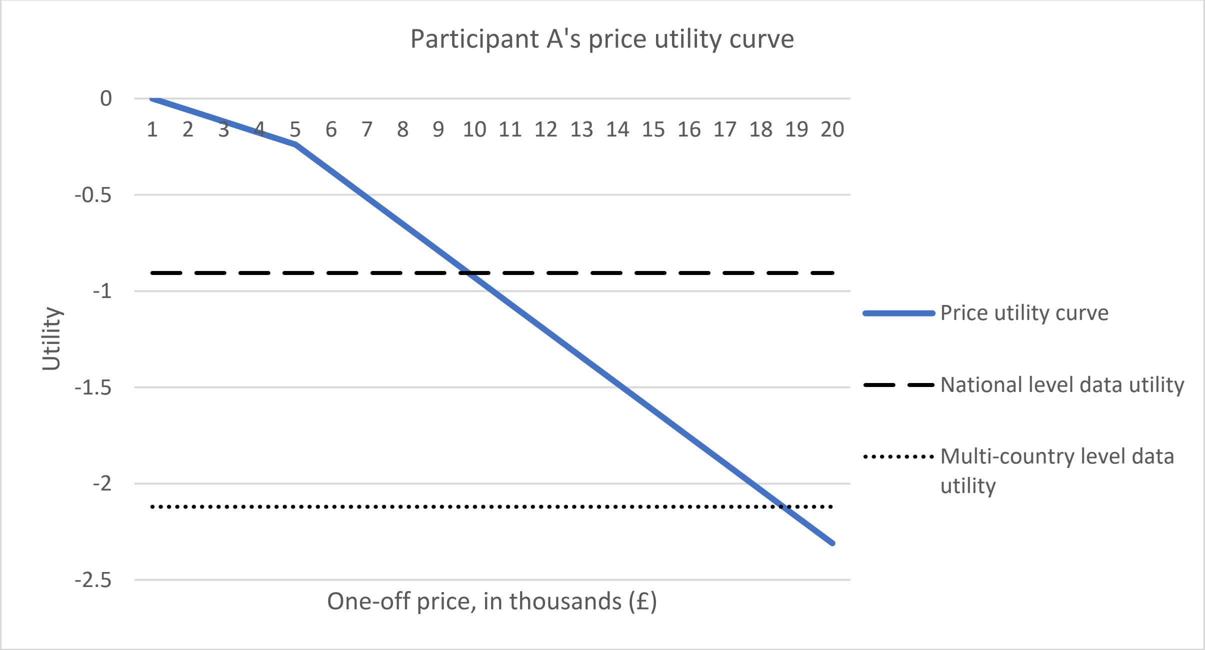
There are two ways of assigning a figure to participants' willingness to pay for attributes and levels. The software outputs estimates of individual level preferences as part of the hierarchical Bayesian multinomial logit model estimates, which are then aggregated to produce the results described above. This also means that it is possible to estimate individual-level marginal rates of substitution and, when price is an attribute, marginal willingness to pay.[5] Taking as an example the marginal willingness to pay for increased data granularity, for a selected participant, A, the software generates a figure of $11,860 as their marginal willingness to pay for the subnational data instead of national data.

For intuition, it is useful to look at a relationship between price and utility for the selected attribute that can be plotted for each individual. An example showing the marginal willingness to pay for different levels of data granularity for participant A is shown in Figure 6. Only the utilities for the three price levels included in the survey have been estimated, and the relationship is assumed to be linear. The solid line maps these utilities of each price level for participant A, with the utility of the lowest price ($1,000) normalised to zero. The utility levels of different levels of granularity are also shown as dashed and dotted lines, with the most granular level of data normalised to zero. These lines do not depend on prices in the analysis we selected. Other forms of discrete choice analysis do allow for interactions between attributes, and therefore for willingness to pay for a new feature to be plotted as a function of price.

At the intersection of these lines, the utility loss of paying that price relative to paying $1,000 is equivalent to the utility loss of reduced granularity relative to the subnational data. Participant A would therefore pay for up to that amount to have access to subnational data over the lower level of granularity. For this participant, having access to national level data instead of the more granular subnational dataset gives a utility of -0.90594. Interpolating from the graph means that Participant A also has a utility of -0.90594 when they are paying $9,831.89. This means that the participant is indifferent between paying $9,931.89 for subnational data, or $1,000 for national level data, giving a marginal willingness to pay for the subnational data instead of national data equal to $8,931.89. As the willingness to pay calculated using the hierarchal Bayesian modelling is $11,860, the lower figure here is likely due to the assumption of linearity rather than convex preferences.

[5] The parameter estimation method is Markov chain Monte Carlo (Medova 2008).

Figure 6: Individual price-utility curve



It is also possible to use individual-level results to understand the heterogeneity within the sample. For example, Figures 7 and 8 show scatter plots of the individual level parameters for different levels of the attributes of timeliness and interoperability. The most frequent and the most interoperable formats were used as baselines, therefore the expected patterns were for parameters to be concentrated in the bottom left quadrant. This is the case in the interoperability graph (Figure 8), but the preference for annual over twice yearly data for many participants means the cluster in the timeliness graph (Figure 7) is further to the right, and suggests that for this data set the attribute is not all that important.

Figure 7: Individual parameters for different levels of timeliness

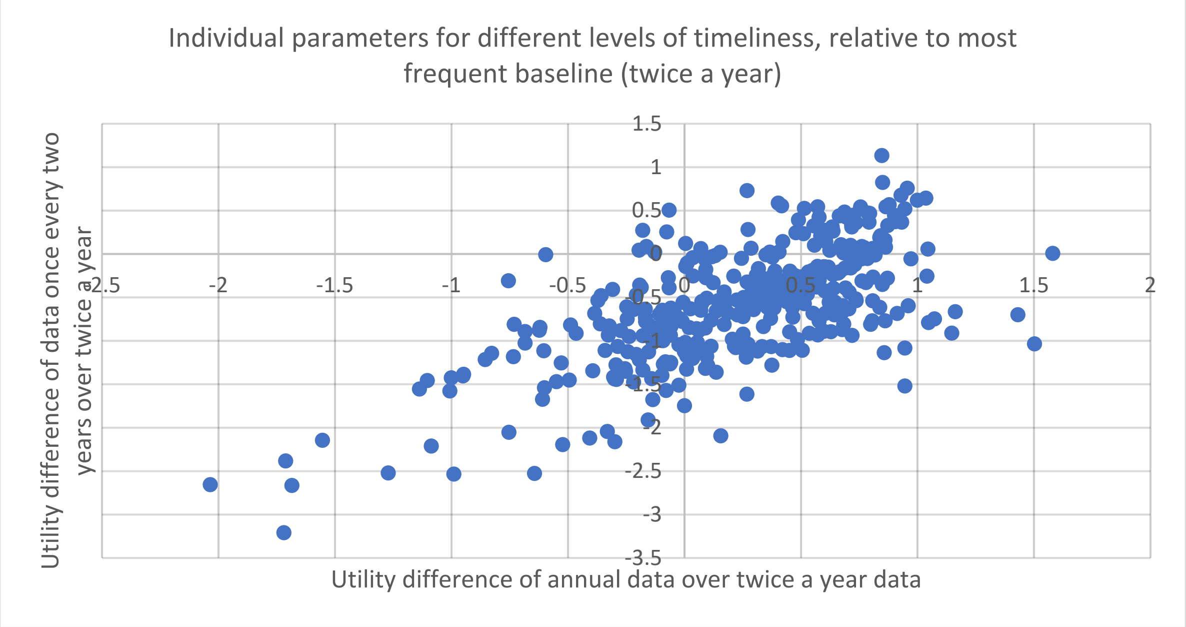
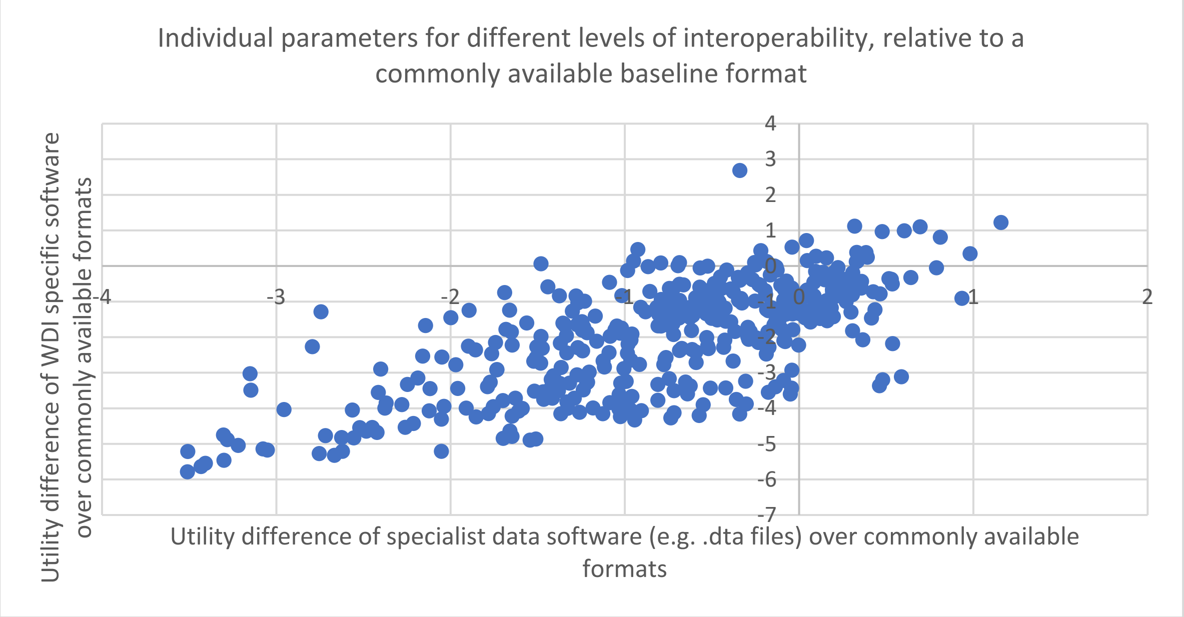


Figure 8: Individual parameters for different levels of interoperability



Finally, out of the 401 respondents, 92 gave further feedback in the open comment box at the end of the survey. Of these, 29 indicated confusion around the methodology or survey design, including ten comments confused about the hypothetical scenario. A further 17 disagreed in principle with the concept of paying for datasets such as the World Development Index, saying they should be free to access and taxpayer funded – an understandable comment albeit one which ignores the well-known challenge of financing public goods and the pressures on public bodies to generate some income. Most of the remaining comments talked about specific attributes (22) or were generic comments about the survey or dataset (16). The other comments relevant to the methodology were mainly criticising the choice of dataset, or the need to specify more context and content description in order to make an informed choice between datasets. The comments indicate that the problems of complexity of method are present even when the sample is limited to a relatively educated, homogeneous, and data-literate group. This is illustrated by the final comments which seemed to indicate some participants believed that the survey indicated a move to create a paywall for the World Development Indicators, despite the care we took to highlight its hypothetical nature in the preamble.

Limitations

One limitation of this methodology is therefore the complexity of the choices facing respondents, such that a certain level of understanding is required for meaningful responses even with a limited number of characteristics. Testing a larger number of attributes would make the survey too long and complex, and so some features of data our open question suggested could be important will have to be omitted. Furthermore, to elicit the total social value of a dataset, the valuations of both experts and nonexperts are required. However, the method requires a good understanding of the different attributes under investigation. There is therefore a trade-off between a more diverse sample with lower quality responses, or less diverse sample with higher quality.

For this experiment, we chose to sample knowledgeable individuals using a dataset which should have been broadly familiar to them. Even so, it is apparent that great care is needed to be taken with survey design to avoid confusion. Surveys need to provide sufficient detail and respondents should be screened for understanding of the scenario they face in their choices. By far the easiest way to convey what information a dataset will contain is by providing an example version of that dataset that the participant can refer to throughout the survey. Screening questions have objectively correct answers that are unlikely to be chosen by chance. This means that in the analysis stage, the respondents can be split into those who have answered correctly and those who have not, with the results taken from the former only.

Another key issue we faced in the survey design was that there is no accepted “standard” way to define the attributes and their levels in the case of data (Louviere et al 2010). This is particularly important as both attribute selection and the levels set affect the outcomes. In particular, the range of levels suggested can affect the proportion of variation in choices that is due to each attribute. For example, as mentioned above, when looking at the timeliness attribute, it would be expected that timeliness would affect decisions more if the choices presented went from daily updates to updates once a decade, than they currently do varying from biannual to biennial updates. Supplementing the method with detailed interviews in advance of running the survey could help structure the selection of attributes and levels.

As noted above, different datasets have different purposes, and those with individual microdata are likely to be useful for a wide range of decisions. One would expect a good deal of heterogeneity in those cases. We consider the method we tested in this paper to be more suitable for standard aggregate data sets. A larger number of survey respondents would be desirable to improve robustness. Finally, the approach shares the well-known limitations of any contingent valuation methodology, including potential hypothetical or strategic bias.

Discussion

Despite these limitations, this discrete choice experiment suggests the method is a practical tool for the question we posed: how should public controllers of certain data sets go about deciding what data to make freely available, whether and what to charge for some types of access, and how much to invest in data gathering and updating. The method can be used to elicit (marginal) willingness to pay for key attributes.

For example, it could be used to set prices under a ‘freemium’ style business model. This model is often used with financial data, where the freely available prices are published 15 minutes behind, while access to real time prices requires payment. A survey of experts is an appropriate method to establish willingness to pay for ‘professional’ rather than public access, in the case of public bodies which – unlike private sector information providers – have social welfare as their objective function and do not have the scope to experiment with prices actually charged. Similarly, the approach could be useful for public services with a wealth of data that could be beneficial for other users but need to cover or justify the costs of investing in a secure and reliable data infrastructure. The figures elicited from a discrete choice experiment need not actually be charged but could be used to estimate a social return on investment in data.

Conclusion

In conclusion, we have shown that a method widely used in other contexts could give holders of public data some insight into the potential social benefit of their data through an application of discrete choice methodology. The surveys and sampling need careful thought, but this approach could help inform decisions about how much to invest in data or how to price which attributes of a data set for commercial sale to cover some of the costs of investment in data gathering and maintenance. Given the acknowledged importance of data in the economy, public data sets are an important part of the national economic infrastructure, so empirical insights into their contribution to social welfare are important. The method we tested certainly has limitations, but on the other hand there is an absence of alternatives, so further experiments would be desirable in addressing “the grand challenge” of understanding the value of data.

References

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. (2020). "The Welfare Effects of Social Media." *American Economic Review*, 110 (3): 629-76.

Arietta Ibarra, I, T Caruso, D Hernadez, E.G. Weyl (2020). An Empirical Study of the Value of Data, in progress.

Birch K, Cochrane D, Ward C. (2021). Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. *Big Data & Society*. Online, May. https://doi.org/10.1177/20539517211017308

Blinder, A.S. (1991). Why Are Prices Sticky? Preliminary Results from an Interview Study. *American Economic Review*, 81: 89-100.

Brynjolfsson, Erik, Avinash Collis, W. Erwin Diewert, Felix Eggers, and Kevin J. Fox. (2020). "Measuring the Impact of Free Goods on Real Household Consumption." *AEA Papers and Proceedings*, 110: 25-30.

Carson, R. T., N. E. Flores, and N. F. Meade,. (2001). “Contingent Valuation: Controversies and Evidence.” *Environmental and Resource Economics,* 19: 173–210. https://doi.org/10.1023/A:1011128332243

Coyle D and S Diepeveen, (2021). Creating and governing social value from data. In progress.

Coyle D & W Li, (2021). The Data Economy: Market Size and Global Trade. ESCoE Discussion Paper 2021-09, https://www.escoe.ac.uk/publications/the-data-economy-market-size-and-global-trade/

Coyle, D. and D. Nguyen. (2020). “Free Goods and Economic Welfare,” ESCoE Discussion Paper 2020-18, https://www.escoe.ac.uk/publications/free-goods-and-economic-welfare-escoe-dp-2020-18/

Coyle, D., Diepeveen, S., Wdowin, J., Tennison, J., & Kay, L. (2019). *The Value of Data: Policy Implications.* Bennett Institute for Public Policy, https://www.bennettinstitute.cam.ac.uk/publications/value-data-policy-implications/

Galperti S, A Levkun & J Perego, (2021). The Value of Data, Working paper https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3866625

Ghorbani & Zou, (2019). What Is Your Data Worth? https://arxiv.org/pdf/1904.02868.pdf

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gurel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. (2020). *Towards efficient data valuation based on the Shapley value*. Paper presented at 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, Naha, Japan.

Ker, D. and E. Mazzini (2020), "Perspectives on the value of data and data flows", *OECD Digital Economy Papers*, No. 299, OECD Publishing, Paris, https://doi.org/10.1787/a2216bc1-en

Koutroumpis, P, Aija Leiponen, Llewellyn D W Thomas, (2020). Markets for data, *Industrial and Corporate Change*, Volume 29, Issue 3, June, Pages 645–660, https://doi.org/10.1093/icc/dtaa002

Louviere, JJ, TN Flynn, RT Carson, (2010). Discrete Choice Experiments Are Not Conjoint Analysis,

Journal of Choice Modelling, Volume 3, Issue 3, Pages 57-72

McFadden, D., & Train, K. (2017) *Contingent Valuation of Environmental Goods: A Comprehensive Critique*, Edward Elgar.

Medova, E. (2008). Bayesian Analysis and Markov Chain Monte Carlo Simulation. In Encyclopedia of Quantitative Risk Analysis and Assessment (eds E.L. Melnick and B.S. Everitt). https://doi.org/10.1002/9780470061596.risk0430

Nakamura A & T Ida (2021), Delineating Zero-Price Markets with Network Effects, Working Paper Graduate School of Economics, Kyoto University, http://www.econ.kyoto-u.ac.jp/dp/papers/e-21-002.pdf

Pei, J. (2020). "A Survey on Data Pricing: from Economics to Data Science," in *IEEE Transactions on Knowledge and Data Engineering*, doi: 10.1109/TKDE.2020.3045927.

UNECE (2018). Recommendations for Promoting, Measuring and Communicating the Value of Official Statistics. https://unece.org/DAM/stats/publications/2018/ECECESSTAT20182.pdf

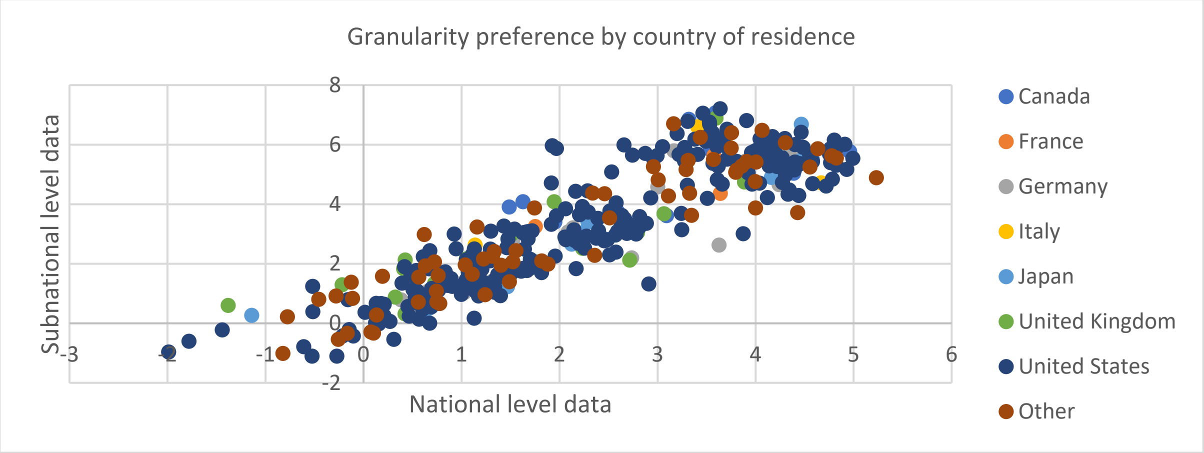
Williams, S. (2021). “Valuing Official Statistics with Conjoint Analysis.” Office of National Statistics

(ONS). https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/valuingofficialstatisticswithconjointanalysisapril2021

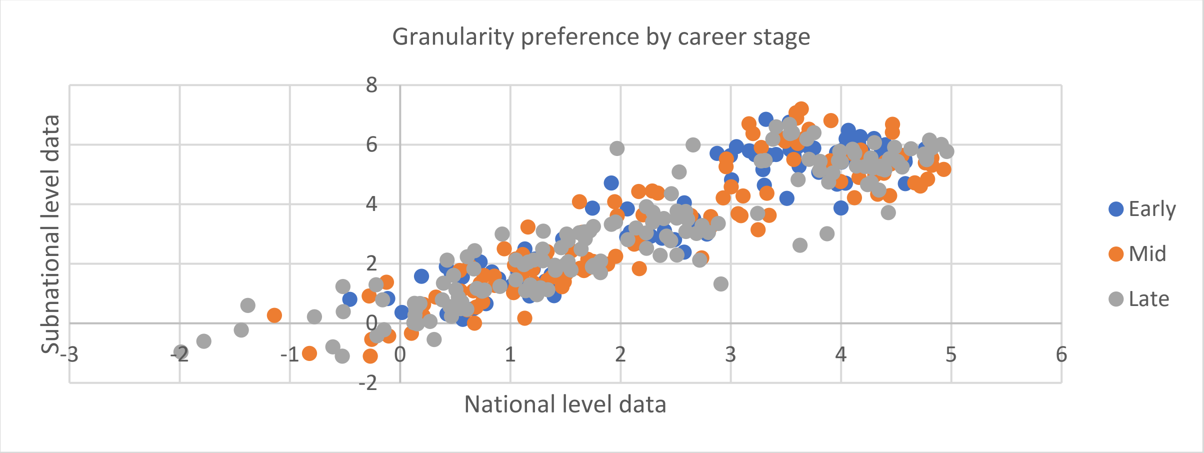
Appendix: individual level preferences

Individual level preferences for granularity by country, stage of career, and prior use of the World Development Indicators[6]

Appendix Figure 1

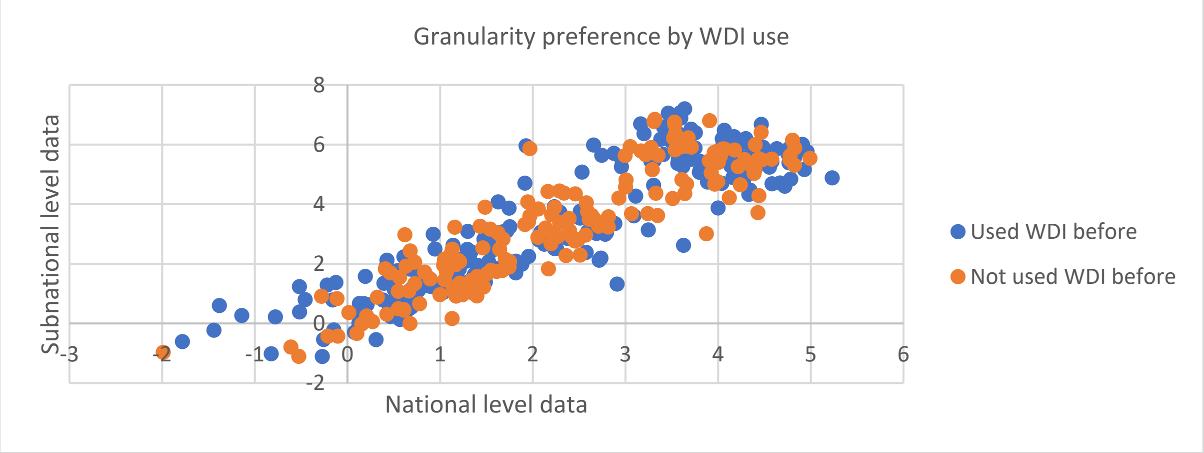


Appendix Figure 2



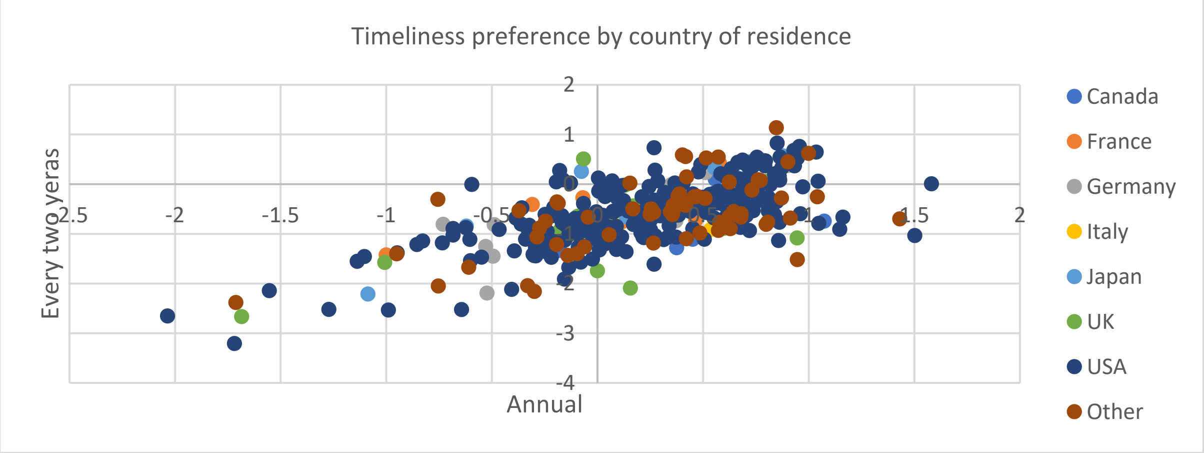
[6] These preferences are relative to the baseline level of grouped country data only. The x-axis value shows the preference of national level and grouped country data over grouped country data only. The y-axis shows the preference of subnational level, national level, and grouped country data over grouped country data only.

Appendix Figure 3

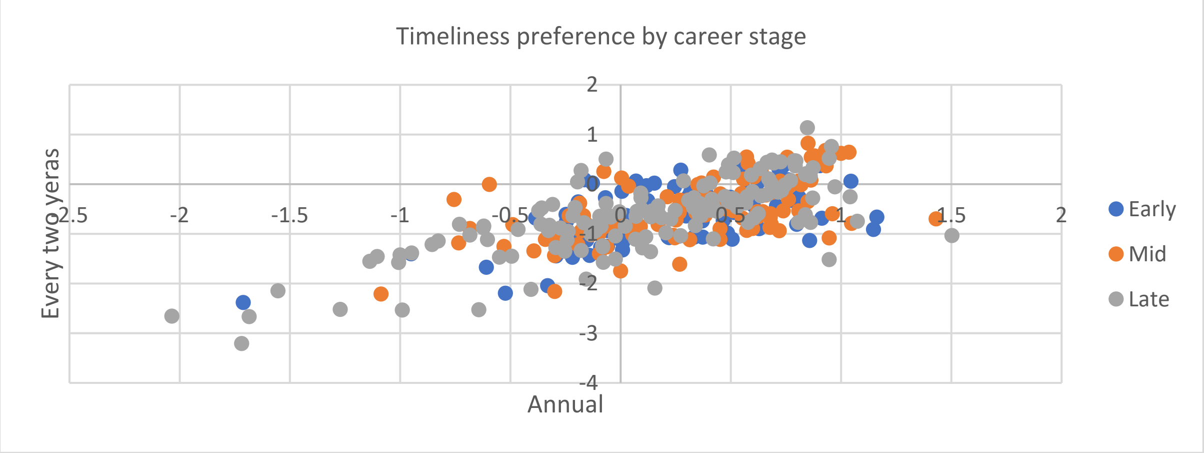


Individual level preferences for timeliness by country, stage of career, and prior use of the World Development Indicators[7]

Appendix Figure 4

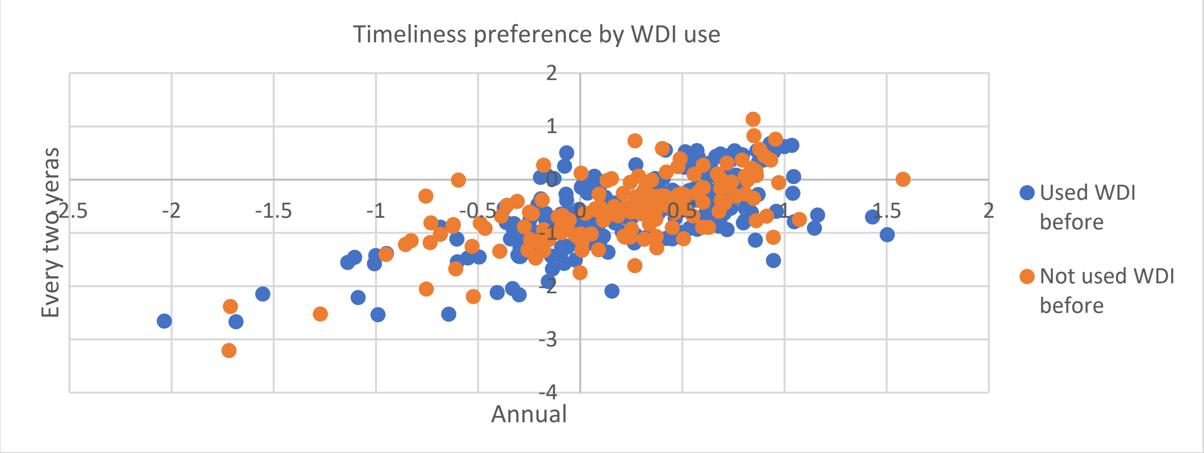


Appendix Figure 5



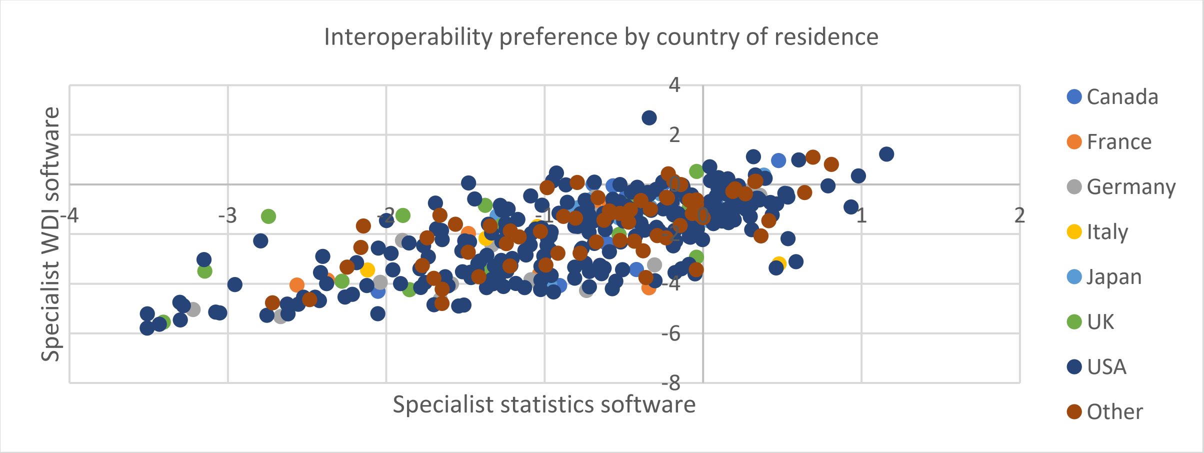
[7] These preferences are relative to the baseline level of data released twice a year. The x-axis value shows the preference of annual data over data twice a year. The y-axis shows the preference of data every two years over twice a year.

Appendix Figure 6

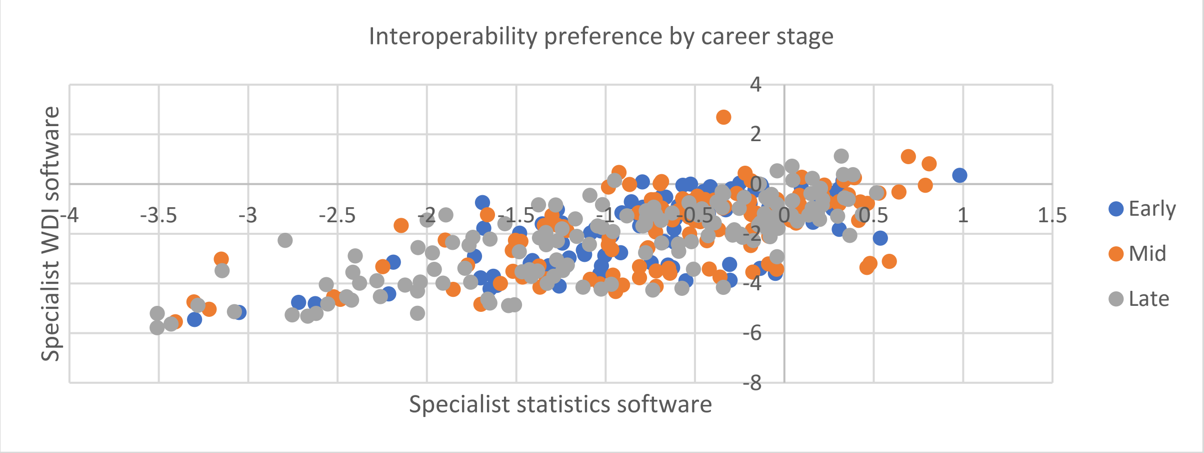


Individual level preferences for interoperability by country, stage of career, and prior use of the World Development Indicators[8]

Appendix Figure 7

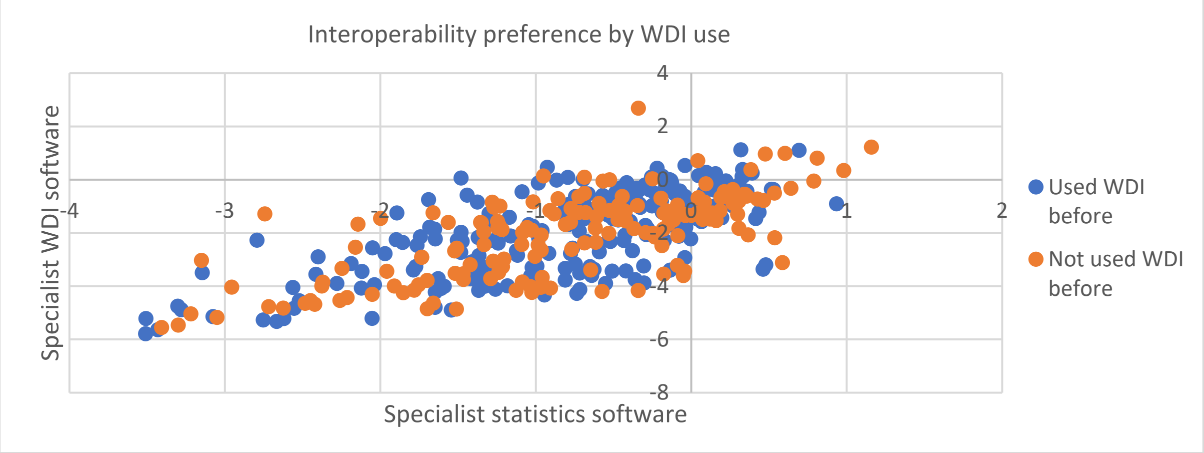


Appendix Figure 8



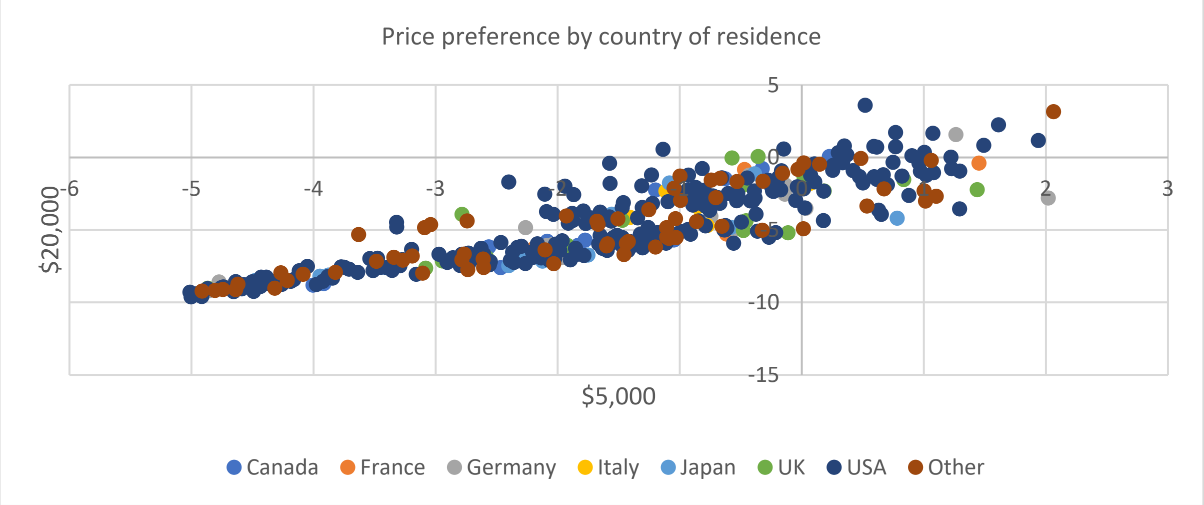
[8] These preferences are relative to the baseline level of data released in a commonly used format (e.g. .xls fils). The x-axis value shows the preference of availability in specialist statistics software (e.g. .dta files) over commonly used formats. The y-axis shows the preference of availability in a hypothetical specialist World Development Indicator software over commonly used formats.

Appendix Figure 8

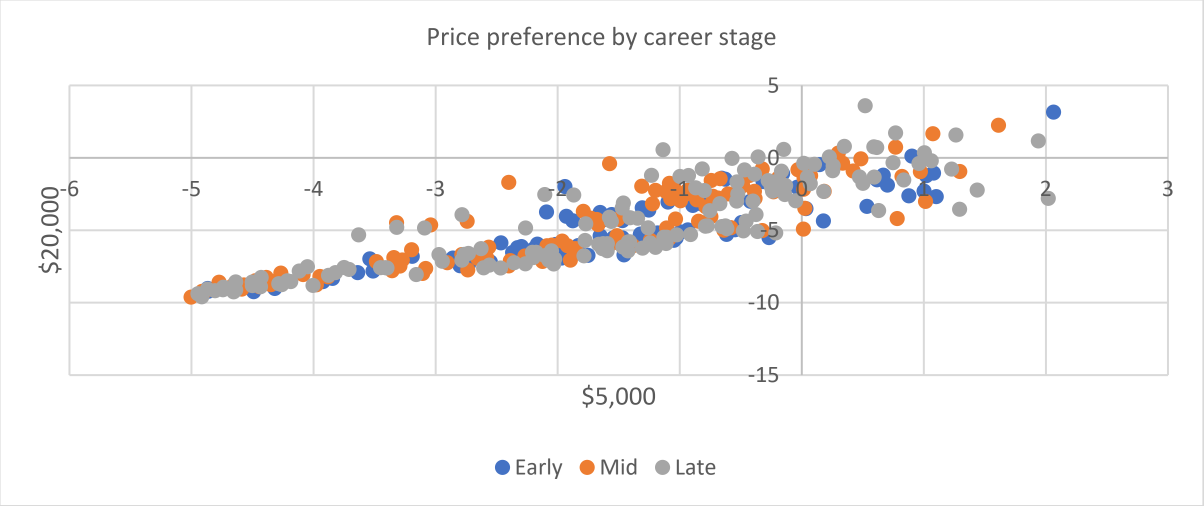


Individual level preferences for interoperability by country, stage of career, and prior use of the World Development Indicators[9]

Appendix Figure 9

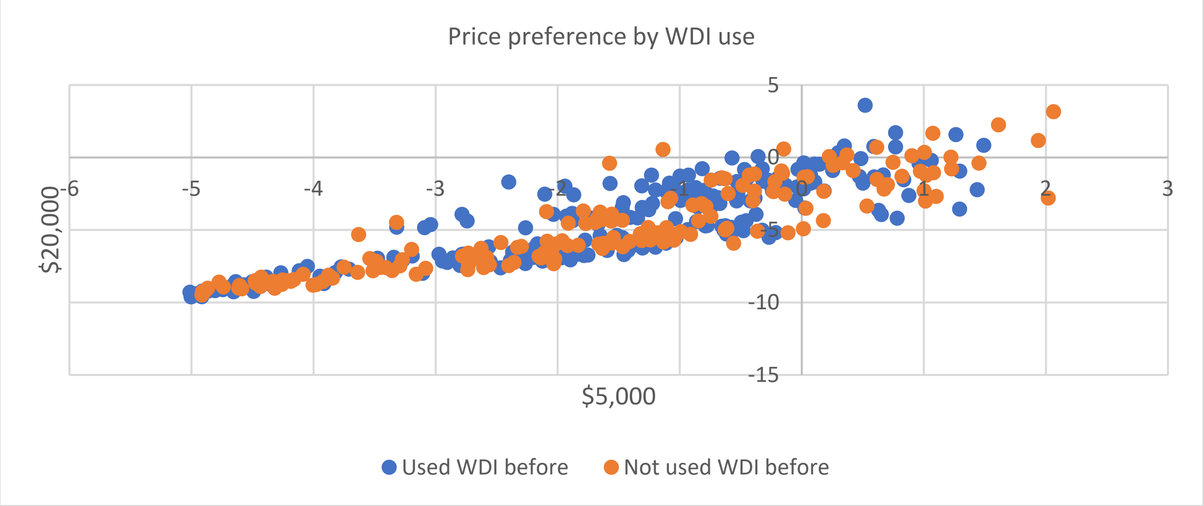


Appendix Figure 10



[9] These preferences are relative to the baseline price of $1,000. The x-axis value shows the preference of paying $5,000 over $1,000 for data. The y-axis shows the preference of paying $20,000 over $1,000 for data.

Appendix Figure 11

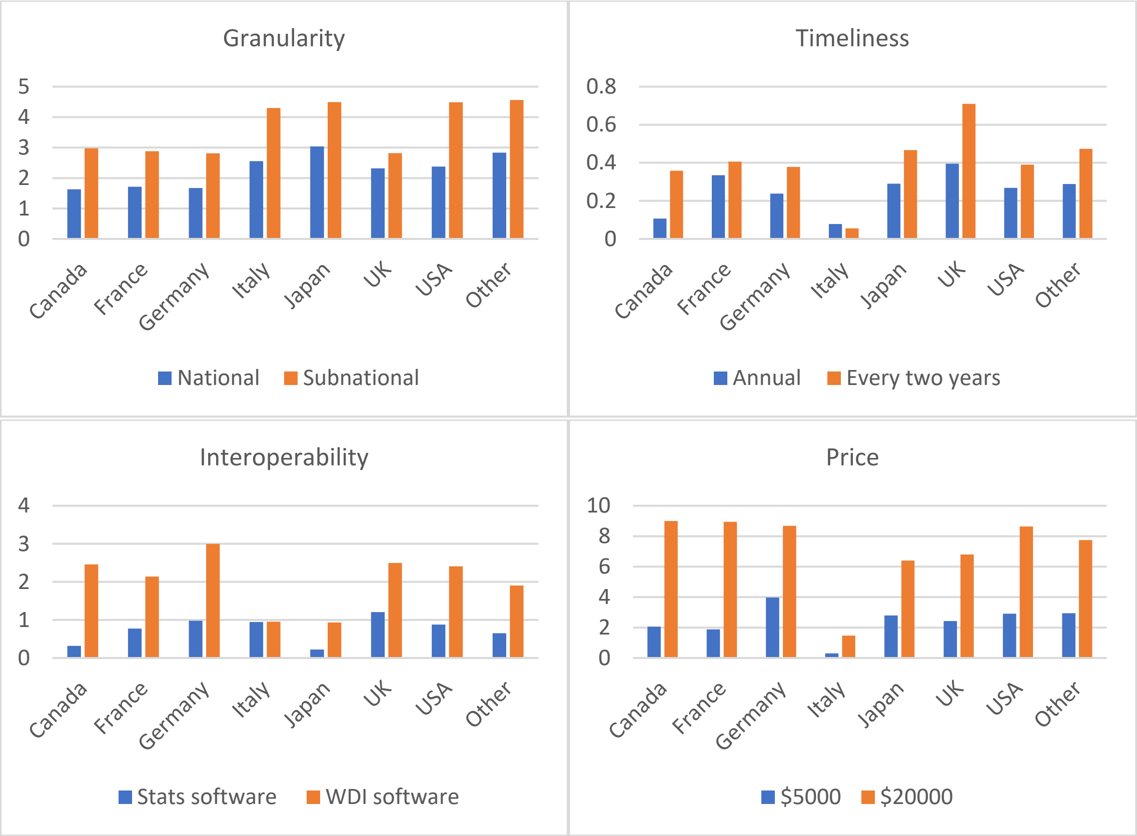


Variation in individual level preferences by country[10]

Appendix Table 1

	Granularity		Timeliness		Interoperability		Price	
	National	Subnational	Annual	Every two years	Stats software	WDI software	$5000	$20000
Canada	1.631913	2.977681	0.106809	0.358163	0.317886	2.459482	2.059585	8.996937
France	1.715597	2.880787	0.334365	0.405836	0.773259	2.14029	1.883494	8.944759
Germany	1.670892	2.814245	0.237949	0.378595	0.979087	2.997379	3.972888	8.686414
Italy	2.556312	4.298559	0.078495	0.055308	0.944113	0.953794	0.306285	1.475268
Japan	3.036604	4.494296	0.290475	0.466926	0.22133	0.932121	2.796442	6.39739
UK	2.315735	2.818969	0.395719	0.709938	1.204903	2.495757	2.424493	6.803129
USA	2.377637	4.491555	0.268132	0.390255	0.876874	2.406952	2.911789	8.638255
Other	2.833836	4.562547	0.288258	0.473737	0.650125	1.902543	2.938468	7.741698

Appendix Figures 12-16



Variation in individual level preferences by career stage

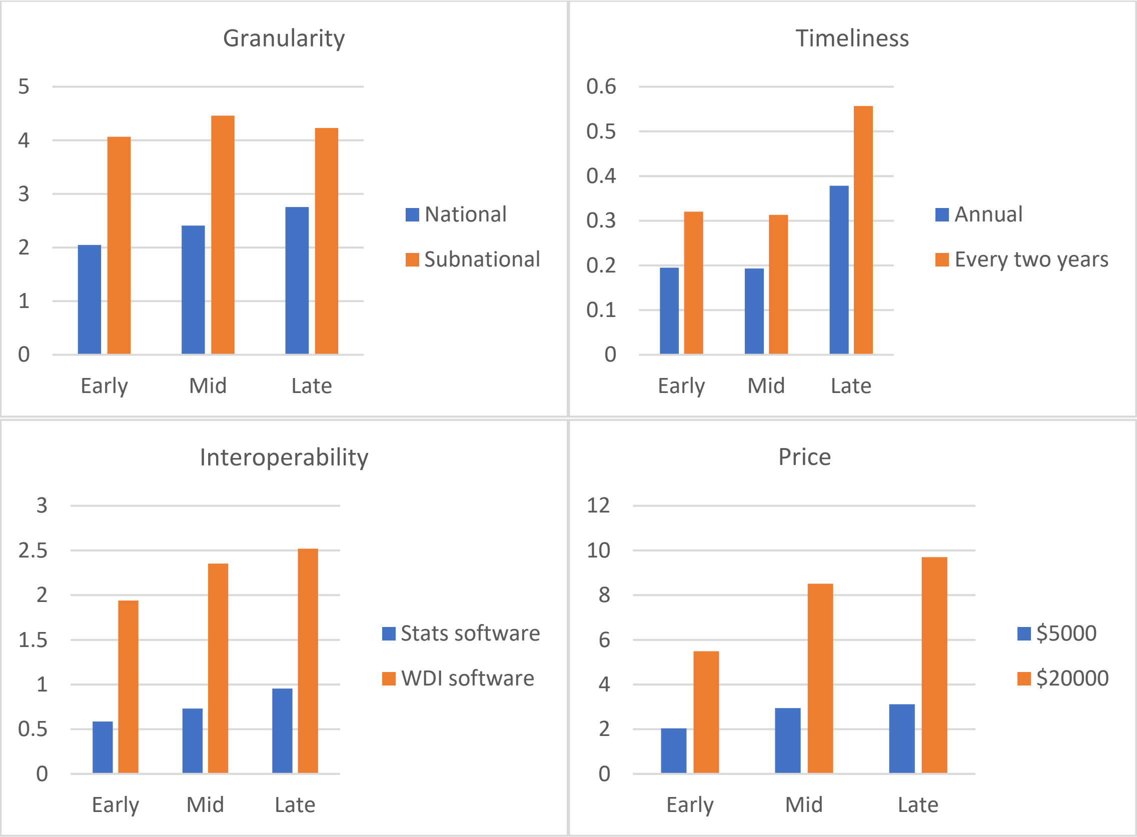
Appendix Table 2

	Granularity		Timeliness		Interoperability		Price	
	National	Subnational	Annual	Every two years	Stats software	WDI software	$5000	$20000
Early	2.050082	4.066457	0.194933	0.320383	0.585908	1.940088	2.032377	5.488351

[10] Please note that some countries (e.g. Italy) have a low n value, and so chance resulting in similar preferences for certain attributes may better explain their low variation than any systematic difference due to country culture.

Mid	2.410566	4.460329	0.193216	0.313122	0.731302	2.352965	2.948104	8.513173
Late	2.757474	4.232105	0.378404	0.556694	0.95416	2.520268	3.117122	9.704492

Appendix Figures 17-20

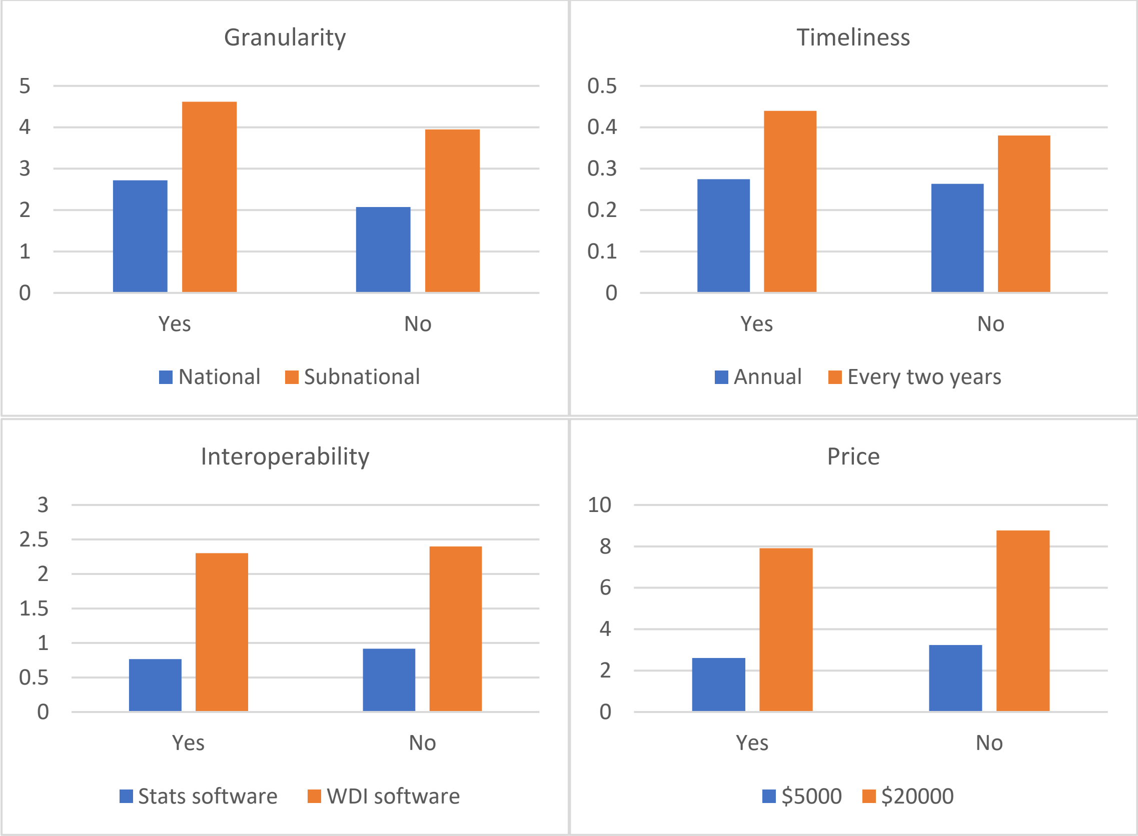


Variation in individual level preferences by prior use of the WDI

Appendix Table 3

	Granularity		Timeliness		Interoperability		Price	
	National	Subnational	Annual	Every two years	Stats software	WDI software	$5000	$20000
Yes	2.721455	4.618994	0.274573	0.43984	0.766813	2.300362	2.612303	7.910703
No	2.072907	3.948561	0.26329	0.38016	0.916369	2.399275	3.231766	8.767759

Appendix Figures 21-24



UNIVERSITY OF
CAMBRIDGE

Bennett Institute
for Public Policy
Cambridge

www.bennettinstitute.cam.ac.uk