Self-identification of occupation in web surveys: requirements for search trees and look-up tables

Survey Methods: Insights from the Field

Kea Tijdens, University of Amsterdam, Amsterdam Institute for Advanced Labour Studies (AIAS), The Netherlands, k.g.tijdens@uva.nl

How to cite this article : Tijdens, K. (2015). Self-identification of occupation in web surveys: requirements for search trees and look-up tables. Survey Insights: Methods from the Field. Retrieved from https://surveyinsights.org/?p=6967

DOI : 10.13094/SMIF-2015-00008

Abstract : Can self-identification of occupation be applied in web surveys by using a look-up table with coded occupational titles, in contrast to other survey modes where an open format question with office-coding has to be applied? This article is among the first to explore this approach, using a random sampled web survey (N=3,224) with a three-level search tree with 1,603 occupations and offering a text box at the bottom of each 3rd level list. 67% of respondents ticked in total 585 occupations, of which 349 by at least two respondents and 236 by only one, pointing to a long tail in the distribution. The text box was used by 32% of respondents, adding 207 occupational titles. Multivariate analysis shows that text box use was related to poor search paths and absent occupations. Search paths for five of the 23 first-level entries should be improved and the look-up table should be extended to 3,000 occupations. In this way, text box use and, thus, expensive manual coding could be reduced substantially. For such large look-up tables semantic matching tools are preferred over search trees to ease respondent's self-identification and thus self-coding.

Introduction

Many surveys have one or more questions with thousands of response categories, the so-called long-list variables such as occupation, industry, car brand, medical drugs, company name and alike. This paper focusses on the measurement of occupations, addressed in most socio-economic and health surveys with a question 'What is your occupation?' or similar (see for the phrasing of this question in more than 30 surveys, Tijdens 2014a).

For long-list survey questions typically an open-ended format is used, followed by expensive and time-consuming coding of the answers after the field work, usually called office-coding or post-survey coding. Alternatively, closed format survey questions could be used, whereby respondents self-identify their occupation. This format however cannot be used for questions with thousands of response categories because in most survey modes the number of response categories is limited. In the CATI mode at most 5 categories can be asked, because otherwise respondents will not memorize. These categories are inevitably highly aggregated. In PAPI the categories shown in a print survey is limited to the maximum number of categories printed on one page, which is around 50. In CAPI it is common to use show cards,

implying the same limits as in PAPI. In CAWI however closed format questions offer new opportunities because CAWI allows for exploiting a look-up table with thousands of response categories. If made available by the survey holder, the CAPI mode also allows for using such a computer-based tool for interviewers to identify respondents' occupations.

In two ways CAWI respondents can self-identify their occupation in a look-up table. First, a search tree or an 'IPod menu' as it is sometimes called allows respondents to navigate through the look-up table by means of a two-level or three-level search tree. Second, semantic matching allows respondents to self-identify their occupation by typing text whereby matches with words in the look-up table are instantly shown. Respondents then select the most relevant match, slightly similar to Google Search. In both ways, the look-up table serves as a prompted survey question, because respondents understand what kind of answers the survey holder is looking for. In the case of occupations, this is advantageous because it prevents responses at various levels of aggregation, thereby avoiding vague occupational titles such as clerk or teacher. Few studies have been conducted regarding the use of search trees and look-up tables in web surveys. Among others Couper et al (2012) conducted a web survey aiming at respondents' self-identification of drugs they used. There is definitely a need to deepen our understanding of respondents' self-identification by means of look-up and how they find their way in these tables. Our final aim is to make suggestions for improvements in search trees and look-up tables for use in web surveys.

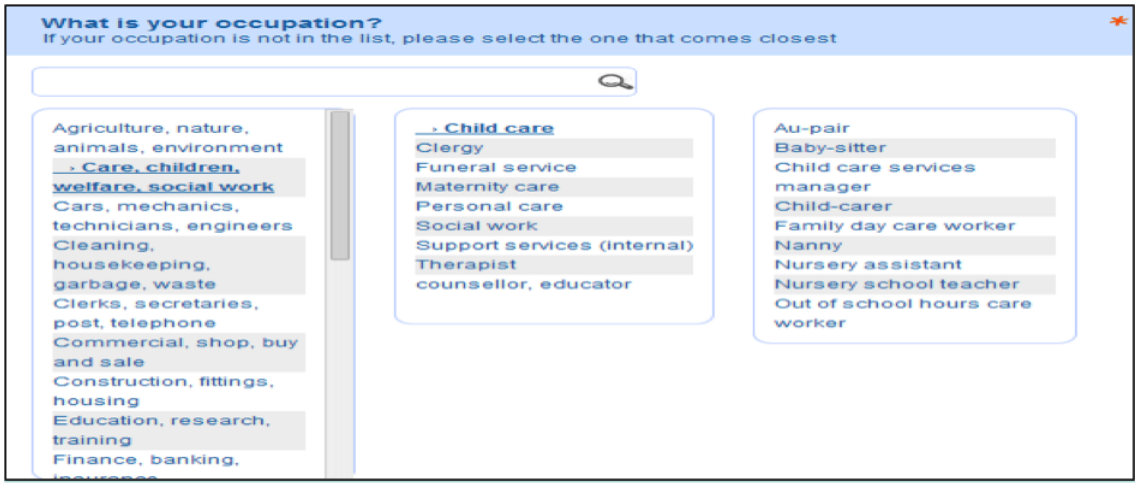
Search trees and semantic matching tools both need a look-up table, but they pose different requirements to such a table. A search tree limits the number of entries in the table because any level in the search tree should preferably not show more than 20 to 25 entries, depending on screen and font size, thus the table is maximized to 8,000 to 10,000 entries. However, maximization of search trees is not advisable because earlier research has shown that the number of characters in the search tree increases the probability of survey drop out (Tijdens 2014b). In contrast, a look-up table for semantic matching should preferably hold as many entries as possible to provide high matching scores. Drafting a limited look-up table for a search tree faces some challenges. First, the stock of occupational titles is very large and may easily exceed the 10,000s. Therefore a limited set should address the most frequent occupations to be effective, though it is difficult to know beforehand which occupations will be most frequently mentioned. Second, the 10,000s job titles are very unequally distributed in national labour forces, depicting a highly skewed distribution with a very long tail of large numbers of rare occupations. In surveys with relative small sample sizes a look-up table will therefore most likely include many occupations that are never selected by respondents. In this article we will explore the frequently-selected occupations versus the hardly-selected ones, the latter reflecting the long tail of the distribution.

For many years the volunteer, continuous WageIndicator web survey on work and wages and the WageIndicator Salary Check have been using a three-level search tree with a look-up table of slightly over 1,600 occupational titles, currently applied in approximately 80 countries. Initially, the web survey only exploited a search tree, but since a few years respondents can choose between the search tree and the semantic matching tool. Figure 1 provides a screen shot of the search tree. The principles underlying this search tree and look-up table, such as the search paths, the alphabetic sorting, the skill levels, the corporate hierarchies, and a range of readability issues, such as the wording of occupations and their translations, have been explained elsewhere (Tijdens, 2010). Note that the search tree does not follow the hierarchy in the International Standard Classification of Occupations (ISCO), because ISCO is designed for classification purposes and not for facilitating respondents' self-identification. Note also that the look-up table does not provide any job description. Respondent's self-identification is therefore solely based on the job title.

Even though millions of web visitors use the Salary Check and hundreds of thousands respond to the web

survey, WageIndicator web managers receive less than one complaint per month about the search tree or the look-up table, hence, the quality of the search tree and its look-up table is assumed to be sufficient to meet the requirements for self-identification for large groups in the labour force. However, when the semantic matching tool was introduced in the web survey the look-up table could not be extended due to budgetary reasons. Mid 2015 a new project has started, allowing for an extension of the look-up table. This stresses the need to explore the quality of the search tree and the look-up table. This paper aims to do so, using data from a representative web survey in the Netherlands.

Figure 1 Three level search tree in the WageIndicator web survey for Great Britain



Source: http://www.paywizard.co.uk/main/pay/salarysurvey/salary-survey-employees, accessed 8 AUG 2014

Data and methods

Data

This paper uses the data of the LISS (Longitudinal Internet Studies for the Social Sciences) panel. LISS is a probability-based online panel in the Netherlands and consists of 5,000 households, comprising of 9,219 individuals aged 16 and over (October 2009). The LISS panel is part of the MESS project (Measurement and Experimentation in the Social Sciences) and it is administered by CentERdata at Tilburg University. The panel was drawn from the population register in collaboration with Statistics Netherlands. Even though the questionnaire is completed online, all people in the sample were recruited in traditional ways by letter, followed by telephone call and/or house visit with an invitation to participate in the panel (for details about the recruitment: Scherpenzeel & Das 2010; Scherpenzeel & Bethlehem, 2011). Households that could not otherwise participate have been provided with a computer and Internet connection.

Each month the panel members are asked to complete a questionnaire. In October 2009 the LISS panel was used for a study to further insight into bias in volunteer samples and to develop methods to adjust for survey bias. The Dutch questionnaire of the WageIndicator web survey was completed by the LISS panel members. Full details of the results of the comparison between the LISS and the WageIndicator data can be found in Steinmetz et al. (2014). Appendix 1 holds the Dutch and the English Codebook of the survey. The current article uses the LISS data regarding the survey question ‘What is your occupation?’ It does not compare the two datasets, but focusses solely on the data from the LISS panel.

In total 5,577 persons responded to this particular LISS survey, reflecting a response rate of 60.5% (Hootsen, 2010). Note that the monthly response of participants varies between 50 and 80%. For our study, only respondents in paid employment were asked about their occupation (3,444 respondents). The occupation question was not asked to students, retired persons and other individuals not active in the labour market. Note that in the LISS panel respondents hardly break off during survey completion because they are instructed not to do so, whereas in the volunteer WageIndicator web survey they do break off to a considerable degree.

The LISS respondents could self-identify their occupation by using a compulsory three-level search tree with a look-up table of 1,603 unique occupational titles, all coded according to the most recent ISCO-08 classification. The occupation search tree used in the LISS panel was similar to the one used in the WageIndicator web survey in the Netherlands. Appendix 2 presents the search tree and its look-up table in Dutch and its translations in English. Note that today the semantic matching technique is widely used for searching an occupation look-up table, particularly by job boards and employment agencies, but in 2009 this technique was not yet in use for the LISS survey. The search tree consisted of 23 entries in level 1 (for example ‘Guards, army, police’), 207 entries in level 2 (for example ‘Guard’), and 1,603 occupational titles in level 3 (for example ‘Bodyguard’ or ‘Doorkeeper’). Hence, the entries in the 3rd level jointly make up the look-up table. In this level, some occupational titles are inserted on more than one place if the search paths were ambiguous, making in total 2,456 entries. To explore the quality of the search tree and its look-up table, the LISS search tree was in one respect different from the WageIndicator tree. On request of the author one extra feature was added. At the bottom of each 3rd level in the search tree an option ‘other’ and a subsequent text box was included, allowing to study to what extent and for which respondents the search tree and its look-up table were not sufficiently detailed.

Research questions

The research questions in our study are threefold. First, what proportion of respondents ticked ‘other’ and reported their occupation via the text box in the search tree? Of these, what proportion could have identified their occupation in the search tree and what proportion had an occupation which was absent in the look-up table?

Second, is the use of ‘other’ and the text box related to the design of the levels in the search tree, to the look-up table or possibly to respondents’ personal characteristics? Here we specify for those who could have identified their occupation and for those whose occupation was absent in the table.

Third, how many of the 1,603 occupational titles were used by the respondents and how often were they ticked? Were the highly skewed distribution and the very long tail reflected in the response and how could this distribution best be described?

Methods

For the purpose of this study, the author coded all text box responses manually and identified whether the coded occupation actually was available in the look-up table or not. Descriptive statistics were used for the first research objective. For the second objective, the likelihood of ticking the text box was modeled for the 23 first level entries and for the personal characteristics age, gender and waged employment. For the third objective the distributional characteristics of the occupation look-up table were used. Note that time stamps or other para-data have not been used in the analyses here.

Results

Use of the search tree and the text box

The first objective aims to present descriptive statistics about the use of the search tree and the text box. Table 1 shows that in the LISS survey drop out during search tree completion is 0.5%, which is much lower compared to the 10 to 20% drop-out rates in the search tree in the volunteer WageIndicator web surveys in Great Britain, Belgium and the Netherlands (Tijdens 2014b). Of the LISS respondents who completed the 3rd level in the search tree, 67% selected an occupation from the look-up table and 32% ticked 'other' and entered their job title in the text box. In a next step the author coded these job titles, using the look-up database with 1,603 job titles. It turned out that 14% could have identified their occupation in the 3rd level of the search tree, but obviously had not found it, implying that they had used different search paths which did not result in their occupation at the 3rd level. Note that the coding process was solely based on the matching of the occupations keyed in with the look-up database, thereby only controlling for typing errors. Job descriptions were not asked in the survey. The remaining 17% expressed an occupation which was indeed absent in the look-up table. If job descriptions had been asked, probably more occupations could have been coded according to the look-up table and had therefore not been classified as absent. Another 0.5% keyed in unidentifiable text.

Table 1 *Distribution over answer categories*

Initial	3444	100%				
Completed level 1 in search tree	3443	100%				
Completed level 2 in search tree	3427	99.5%				
Completed level 3 in search tree	3426	99.5%				
... of which selected an occupation			2313	67.2%		
... of which ticked 'other' and used text box			1113	32.3%		
... ... of which could have found their occupation					497	14.4%
... ... of which occupation was absent					600	17.4%
... ... of which no relevant text					16	0.5%

Source: WageIndicator Questionnaire administered to the LISS panel, October 2009

The 23 first level entries in the search tree are the largest hurdle for respondents, because the number of characters to be read is large and might therefore be difficult to comprehend. Per entry Table 2 shows the distribution over the three groups – the respondents who selected an occupation in the search tree, the ones who ticked the text box but could have found their occupation, and the ones who ticked the text box and the occupation was absent. Table 2 points to the most problematic entries. In the first level entry 'Oil, gas, mining, utilities' only 18% of respondents could identify their occupation in the look-up table, whereas 56% keyed in an occupational title that was absent in the table, pointing to shortcomings in the look-up table. This shortcoming also appears to be the case for the entry 'Cars, mechanics, technicians, engineers' where the share of absent occupations is high with 30%. For the entry 'Food manufacturing' only 44% of respondents could identify their occupation. Here the problem seems to be related to shortcomings in the search paths, because 28% respondents could not find their occupations although these were included.

The last column in Table 2 presents the ratio between the columns C and D. Higher ratios point to relative difficulties in the look-up table, whereas lower ratios do so for the search paths. For the entry 'Management, direction' the look-up table obviously has many missing occupations, whereas for the entry 'Clerks, secretaries, post, telephone' the search paths need to be improved.

Table 2 Distribution over answer categories, breakdown by first level entries

	Completed level 3 in search tree &			Total	
	Completed level 3 in search tree and ...			Total	Ratio
First level of search tree	selected an occu-pation (B)	ticked other but could have found occupation (C)	ticked other and occupation was absent (D)		Col D/C
Transport, logistics, port, airport	74.2%	11.6%	14.2%	155	1.2
Management, direction	72.6%	6.8%	20.5%	73	3.0
Oil, gas, mining, utilities	18.5%	25.9%	55.6%	27	2.1
Media, graphic, printing, culture, design	58.1%	20.9%	20.9%	86	1.0
Marketing, PR, advertising	62.2%	20.0%	17.8%	45	0.9
Legal, administration, inspection, policy adviser	55.2%	24.8%	20.0%	145	0.8
Language, library, archive, museum	62.5%	16.7%	20.8%	24	1.3
IT, automation, telecommunication	61.5%	16.8%	21.7%	143	1.3
Industrial production, manufacture, metal	56.4%	17.6%	26.1%	188	1.5
HRM, labour intermediary, organisation	68.8%	12.5%	18.8%	48	1.5
Hospitality, tourism, leisure, sports	73.9%	13.6%	12.5%	176	0.9
Health care, paramedics, laboratory	73.9%	10.4%	15.7%	383	1.5
Guards, army, police	81.2%	8.7%	10.1%	69	1.2
Food manufacturing	43.6%	28.2%	28.2%	39	1.0
Finance, banking, insurance	72.7%	12.7%	14.6%	205	1.2
Education, research, training	70.5%	12.8%	16.7%	312	1.3
Construction, fittings, housing	64.3%	20.4%	15.3%	157	0.8
Commercial, shop, buy and sale	73.4%	11.7%	14.9%	349	1.3
Clerks, secretaries, post, telephone	80.7%	12.3%	7.1%	212	0.6
Cleaning, housekeeping, garbage, waste	78.8%	12.1%	9.1%	99	0.8
Cars, mechanics, technicians, engineers	53.2%	15.6%	31.2%	109	2.0
Care, children, welfare, social work	61.5%	16.7%	21.8%	275	1.3
Agriculture, nature, animals, environment	58.0%	17.0%	25.0%	88	1.5

Source: WageIndicator Questionnaire administered to the LISS panel, October 2009, excluding respondents who did not complete all three levels (18 observations), who entered unidentifiable text (16 observations) and who had missing values for gender (3 observations)

Who uses the text box?

The second objective is to explore whether the use of the text box is related to the design of the search tree or to respondents' personal characteristics. Model 1 in Table 3 explores this for those respondents who could have identified their occupation versus those who selected an occupation in the search tree, hence identifying problematic search paths (496 versus 2,312 respondents). In Model 2 we do so for those whose occupation is absent in the database versus those who selected an occupation in the search tree, hence pointing to problems in the look-up table (599 versus 2,312 respondents).

For respondents selecting the first level entry ‘Oil, gas, mining, utilities’ the odds ratios in Model 1 increases approximately 9 times compared to the reference entry. For respondents who selected the first level entry ‘Food manufacturing’ the odds ratios increases 4 times. The first level entries ‘Media, graphic, printing, culture, design’, ‘Legal, administration, inspection, policy adviser’, and ‘Construction, fittings, housing’ reveal increases of more than 2 times. The effects of the search tree entries hardly change once personal characteristics are included in Model 1b.

In Model 2 – entering a job title that is absent in the look-up table -, the odds ratios for the first level entry ‘Oil, gas, mining, utilities’ increase even 15 times compared to the reference entry. For two entries the odds ratio increase 3 times, namely ‘Food manufacturing’ and ‘Cars, mechanics, technicians, engineers’. For another two entries they increase more than 2 times, namely ‘Industrial production, manufacture, metal’ and ‘Agriculture, nature, animals, environment’. Here too the effects hardly change when personal characteristics are entered into Model 2b.

Concerning the personal characteristics Table 3 shows that for respondents in waged employment the odds ratio in Model 1b decreases with 26%, whereas the odds ratio in Model 2b is not affected significantly. For women, the odds ratio increases with 40% in Model 1b whereas the odds ratio in Model 2b is not affected significantly. In both models, the odds ratios increase with age.

In conclusion, the use of the text box is highly affected by the design of the search tree. In particular, five of the 23 first level entries point to ambiguous search paths, requiring that more occupations from the look-up table are to be inserted in these entries. Another five first level entries point to absent occupations and these have to be added to the look-up table.

Table 3 Odds ratios and standard errors of respondents' probabilities of using the text box although occupation is present in search tree (Model 1a and b) and of using the text box with occupation absent in search tree (Model 2a and b), both versus selecting an occupation in the search tree

	M1a	M1b	M2a	M2b
	Exp(B)	Exp(B)	Exp(B)	Exp(B)
Management, direction	0.603	0.586	1.479	1.441
	(0.532)	(0.533)	(0.374)	(0.375)
Oil, gas, mining, utilities	8.944***	9.497***	15.682***	16.012***
	(0.638)	(0.641)	(0.566)	(1.441)
Media, graphic, printing, culture, design	2.300**	2.012*	1.882*	2.014*
	(0.374)	(0.378)	(0.36)	(16.012)
Marketing, PR, advertising	2.054	1.873	1.494	1.69
	(0.459)	(0.465)	(0.464)	(2.014)
Legal, administration, inspection, policy adviser	2.875***	2.581***	1.895**	1.989**
	(0.323)	(0.326)	(0.318)	(1.69)
Language, library, archive, museum	1.704	1.187	1.742	1.862
	(0.617)	(0.626)	(0.566)	(1.989)
IT, automation, telecommunication	1.742	1.835*	1.841*	1.88**
	(0.342)	(0.344)	(0.313)	(1.862)
Industrial production, manufacture, metal	1.989**	2.105**	2.416***	2.372***
	(0.322)	(0.324)	(0.29)	(1.88)
HRM, labour intermediary, organisation	1.162	0.992	1.426	1.616
	(0.511)	(0.517)	(0.442)	(2.372)
Hospitality, tourism, leisure, sports	1.179	1.075	0.885	1.086
	(0.337)	(0.346)	(0.328)	(1.616)
Health care, paramedics, laboratory	0.903	0.750	1.108	1.226
	(0.305)	(0.314)	(0.273)	(1.086)
Guards, army, police	0.685	0.747	0.653	0.66
	(0.499)	(0.500)	(0.464)	(1.226)
Food manufacturing	4.134***	4.134***	3.382***	3.451***
	(0.463)	(0.465)	(0.452)	(0.66)
Finance, banking, insurance	1.115	1.068	1.052	1.108
	(0.331)	(0.333)	(0.307)	(3.451)
Education, research, training	1.162	1.028	1.236	1.318
	(0.306)	(0.311)	(0.279)	(1.108)
Construction, fittings, housing	2.024**	2.196**	1.242	1.245
	(0.325)	(0.327)	(0.325)	(1.318)
Commercial, shop, buy and sale	1.023	0.886	1.062	1.239
	(0.304)	(0.311)	(0.278)	(1.245)
Clerks, secretaries, post, telephone	0.971	0.790	0.459**	0.516*
	(0.329)	(0.338)	(0.356)	(1.239)
Cleaning, housekeeping, garbage, waste	0.983	0.764	0.603	0.683
	(0.401)	(0.409)	(0.422)	(0.516)
Cars, mechanics, technicians, engineers	1.873*	1.992*	3.064***	3.092***
	(0.375)	(0.376)	(0.317)	(0.683)
Care, children, welfare, social work	1.739*	1.416	1.856**	2.117***
	(0.303)	(0.312)	(0.277)	(3.092)
Agriculture, nature, animals, environment	1.879	1.648	2.255**	2.273**
	(0.388)	(0.392)	(0.345)	(2.117)
In waged employment (versus self-employed)		0.742**		1.08
		(0.123)		(2.273)
Female		1.401***		0.851
		(0.121)		(1.08)
Age (16-64)		1.010**		1.009**
		(0.004)		(0.851)
Constant	0.157***	0.120***	0.191***	0.124***
	(0.253)	(0.318)	(0.233)	(1.009)
Chi-square (df=22,df=25)	71.21	87.59	116.73	126.43
-2 Log likelihood	2547.29	2530.91	2892.84	2883.13
N	2808	2808	2927	2927

Note: *Reference category Transport, logistics, port, airport;*
*** $p<0.01$, ** $p<0.05$; * $p<0.10$*
Source: *WageIndicator Questionnaire administered to the LISS panel, October 2009.*

Which occupations are ticked?

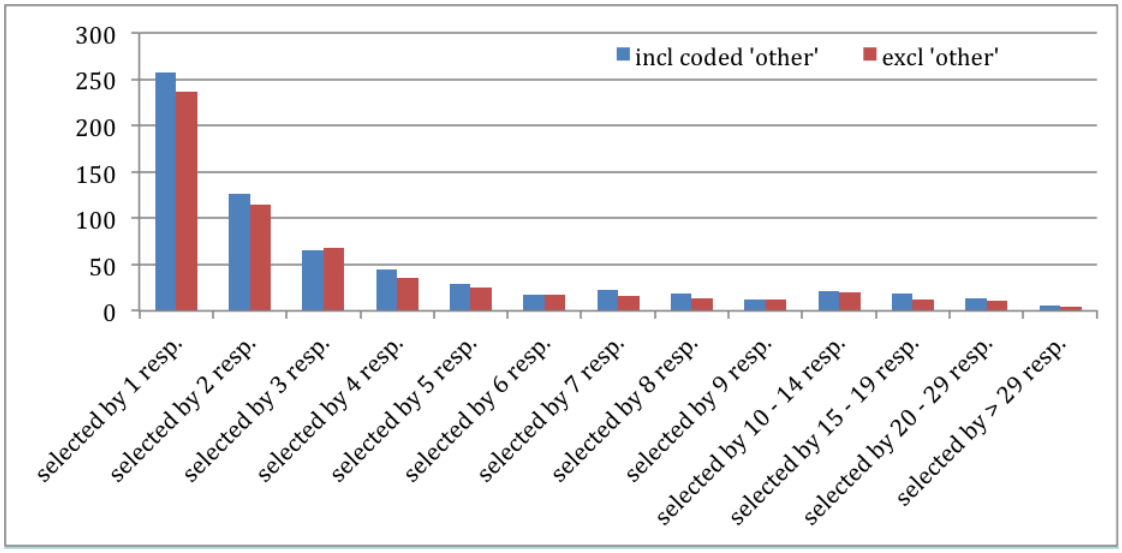
The third objective is to explore how many of the 1,603 occupational titles in the look-up table are used

by the respondents and how often each title is selected. In other words: how is the highly skewed occupational distribution in the labour force reflected in the survey response and how could this distribution best be described?

The 2,313 respondents who did select an occupation in the search tree used 585 of the 1,603 titles in the look-up table. The 497 respondents who completed the text box but could have identified their occupational title used 207 titles from the table, of which 139 were also selected by the group of 2,313 respondents. Jointly these two groups of 2,810 respondents ticked 653 titles.

Figure 2 reveals that very few occupations are selected by 30 respondents or more. This applies to 4 occupations, selected by 10.8% of the 2,313 respondents, who selected an occupation in the search tree. It applies to 6 occupations, selected by 13.0% of the 2,810 respondents, after the text box answers were coded. Frequently mentioned occupations are 'Office clerk', 'Primary school teacher', 'Health associate professional' and 'Elderly aide'. Only 48 respectively 60 titles were selected by 10 to 29 respondents and another 187 respectively 210 titles by 3 to 9 respondents, totaling to 69% respectively 70% of respondents. Another 114 respectively 126 titles were selected by only 2 respondents and 236 respectively 257 occupations only once. In total 1,018 respectively 950 of the 1,603 titles in the look-up table were not selected in our survey.

Figure 2 The number of occupations selected by the 2,313 respondents who selected an occupation and by the 2,810 respondents including those with coded occupation



Source: WageIndicator Questionnaire administered to the LISS panel, October 2009

In total 600 respondents completed the text box and could not have identified their job titles, because they were absent in the look-up table. After cleaning for misspellings and harmonization of gendered job titles, they keyed in 555 different job titles.

In summary, the look-up table included 1,603 occupational titles, of which 653 were selected by respondents (41%). An additional 555 occupational titles should have been listed, but were absent. Assuming that the unlisted occupational titles hold the same rate of 41%, the look-up table should have been extended with 1,362 titles to meet the demands of the LISS sample. Hence, a sample size of 3,444 respondents requires a database of at least 1,603 + 1,362 = 2,965 titles. Larger sample sizes require larger databases.

Conclusion/Discussion

Web surveys allow for respondents' self-coding by using a search tree with a look-up table for the survey question 'What is your occupation?'. This is in contrast to other survey modes which can apply mainly an open format question with office-coding. Using a representative sample of 3,444 web survey respondents in the Netherlands and a three-level search tree with a look-up table with 1,603 occupational titles, 67% of respondents selected an occupation from this table and 32% used the text box, which was included as the last entry at each 3rd level of the search tree. After coding these responses, it turned out that almost half of them could have identified their occupation but had not found and thus pointing to poor search paths, whereas slightly over half expressed an occupation which was absent in the look-up table. Using multivariate analyses, we identified that for five of the 23 search tree's first level entries the likelihood of text box use due to poor search paths was substantially higher than in the remaining entries. For another five entries, of which two were overlapping, the likelihood of an absent occupation was substantially higher. For this reason particularly respondents with an occupation in the entry 'Oil, gas, mining, and utilities' encountered difficulties in identifying their occupation. Older respondents had more difficulties in identifying their occupation in the look-up table, but we have no evidence whether older respondents have more cognitive difficulties in doing so or that the table included fewer occupations associated with older workers.

Given the 10,000s of occupational titles in any national labour force and the long tail of the distribution of workers over occupations, it is not surprising that the skewed distribution was noticed in our sample too. 11% of respondents selected only 4 occupations, 69% selected 235 occupations, and the remaining 20% selected 350 titles. We computed that our look-up table would have needed at least 2,965 instead of 1,603 occupational titles to allow all respondents to select an occupation. Larger samples will need larger tables. Of course, a major challenge relates to identifying which occupational titles should be included in the look-up table, because this has to be determined before a web survey starts. Otherwise, manual coding remains necessary, and this is particularly expensive for the many occupations with relative few jobholders in the long tail. As explained, the number of entries in a search tree is maximized. Therefore, a semantic matching tool should be preferred over a search tree when large numbers of entries are included in the look-up table.

Although beyond the scope of this article, we want to make a comment concerning semantic matching used for respondents' self-identification compared to auto-coders used for office-coding. Although both require a list of coded occupational titles, a look-up table for self-identification is different from a training set for auto coders facilitating machine learning algorithms. During survey completion a semantic matching tool provides respondents with a list of matched occupations while they type a few characters and the match list is adapted when respondents enter more characters. Auto-coders are used after survey completion and they also need to correct text strings for typing errors, for highly aggregated occupational titles, for synonyms, for female/male expressions of occupations, and alike, leading to a set of so-called hard codes. Semantic matching lists will not include hard codes, because typing errors will not lead to matches and respondents will understand instantly that they need to correct for errors for the purpose of a match. Look-up tables for semantic matching should not include the aggregate occupational title 'clerk', but a list of specified clerk titles, inviting respondents to tick one of them. In many industrialised countries, auto-coding has gradually developed, as for example publications by Hoffmeyer-Zlotnik, Hess, Geis (2006) and Bethmann et al (2014) show for Germany.

Survey holders can of course use the search tree and look-up table used in this study (see Appendix 2 for the table in Dutch and in English). The first four digits in the first column reflect the ISCO-08 code. On

request the author can supply translations in other languages. As of mid-2015 an extension of the look-up table is scheduled. The table will be made available by means of an API (Application Program Interface), such that any survey holder can include a link in his/her web survey that calls for this API for the survey question ‘What is your occupation?’.

Appendix 1: The WageIndicator Questionnaire administered to the LISS panel

Dutch version

English version

Appendix 2: The look-up table and its search tree.

WageIndicator occupation database

References

1. Bethmann, A., Schierholz, M., Wenzig, K., Zielonka, M. (2014) *Automatic Coding of Occupations Using Machine Learning Algorithms for Occupation Coding in Several German Panel Surveys.* Presentation VI European Congress of Methodology Utrecht University 24 July 2014
2. Couper, M. P., Zhang, C., Conrad, F. G., Tourangeau, R. (2012) *Database Lookup in Web Surveys.* Paper presented at the 6th Internet Survey Metodology Workshop, Ljubljana.
3. Hoffmeyer-Zlotnik, J. H. P., Hess, D., Geis, A. J. (2006) Computerunterstützte Verco¬dung der International Standard Classification of Occupations (ISCO-88): Vorstellen eines Instruments. *ZUMA-Nachrichten 58/2006*: 101-113.
4. Hootsen, J. (2010) WageIndicator Questionnaire administered to the LISS panel. Tilburg, CentERdata [the dataset is available on request at http://www.centerdata.nl/en/databank/liss-panel-data-0]
5. Scherpenzeel, A. & Das, M. (2010). True longitudinal and probability-based Internet panels: Evidence from the Netherlands. In Das, M., Ester, P. & Kaczmirek, L. (eds.). *Social and Behavioral Research and the Internet: Advances in Applied Methods and New Research Strategies*. Boca Raton: Taylor & Francis.
6. Scherpenzeel, A. & Bethlehem, J. (2011). How representative are online panels? Problems of coverage and selection and possible solutions. In: M. Das, P. Ester and L. Kaczmirek (eds.), *Social Research and the Internet: Advances in Applied Methods and New Research Strategies*. Boca Raton: Taylor & Francis, pp. 105-132.
7. Steinmetz, S., Bianchi, A., Tijdens, K.G., Biffignandi, S. (2014). Improving web survey quality – Potentials and constraints of propensity score adjustments. In: Callegaro, M., Baker, R., Bethlehem, J., Goritz, A., Krosnick, J., Lavrakas, P. (eds.) *Online Panel Research: A Data Quality Perspective*. Chichester: Wiley, pp 273-298.
8. Tijdens KG (2010) *Measuring occupations in web surveys the WISCO database of occupations*. Amsterdam: University of Amsterdam, AIAS Working Paper 10-86. http://www.uva-aias.net/uploaded_files/ publications/WP86-Tijdens.pdf
9. Tijdens, K.G. (2014a). *Reviewing the measurement and comparison of occupations across Europe.* Amsterdam: University of Amsterdam, AIAS Working Paper 149.
10. Tijdens, K.G. (2014b). Drop-out rates during completion of an occupation search tree in web-surveys. *Journal of Official Statistics*, 30 (1), 23–43.