Dropout rates during completion of an occupation search tree in web-surveys

Kea Tijdens, University of Amsterdam / Amsterdam Institute for Advanced Labour Studies (AIAS)

DRAFT!! Not to be quoted without permission of the author, March 24, 2011

Contact: k.g.tijdens@uva.nl

Acknowledgements: This paper builds on research work done for the EU-funded FP6 project EurOccupations (no 028987, 2006-2009, www.eurooccupations.org) and for the WageIndicator web-survey on work and wages (www.WageIndicator.org).

Table of contents

Abstract ........................................................................................................................................................................... 2
1 Introduction .......................................................................................................................................................... 2
2 Reviewing the measurement of occupations ................................................................................................... 3
  2.1 The open and closed response formats in PAPI, CATI, CAPI and CAWI ...................................... 3
  2.2 Three response formats in web-surveys (CAWI) ................................................................................ 4
3 The occupation database for search trees ........................................................................................................ 5
  3.1 Introducing the WISCO database ............................................................................................................ 5
  3.2 Efficient search and readability ................................................................................................................. 5
  3.3 The aggregation level of occupations in the database ........................................................................... 6
  3.4 The skill level of occupations in the database ......................................................................................... 6
  3.5 The corporate hierarchy of occupations in the database ...................................................................... 6
4 Dropout rates and completion time .................................................................................................................. 7
  4.1 Hypothesis, data and methodology ........................................................................................................ 7
  4.2 Dropout rates during search tree completion ........................................................................................ 8
  4.3 Time needed to complete the search tree ............................................................................................... 9
  4.4 Analyses of the response to the open response format question......................................................12
5 Database requirements for text string matching ...........................................................................................13
6 Conclusions .........................................................................................................................................................14
References.....................................................................................................................................................................15
Dropout rates during completion of an occupation search tree in web-surveys

Abstract

Occupation is key in socio-economic research. Web-surveys are disadvantageous because absent interviewers allow unidentifiable or aggregated responses. To facilitate respondent’s self-identification, the multi-country WageIndicator web-survey employs a 3-tier search tree with 1,700 occupations. Using the 2010q2 WageIndicator data for UK, Belgium and Netherlands (24,811 observations), this paper investigates search tree dropout rates, completion time and respondents’ use of an open question following the search tree. Hypotheses relate dropout and completion time to survey burden (number of characters in search path), question adequacy (do respondents have a job title) and respondents’ cognitive capacities. Tier 1 dropout is high and explained by adequacy, tier 3 dropout by burden. Tier 2 and 3 completion time depend on adequacy, burden, capacities and problems finding the right search path. One in five respondents use the open question. 1.7% of the ticked occupations are wrong, pointing to problematic search paths. Requirements for text string matching, a promising alternative, are explored.

Keywords: job title; CAWI; occupation database; ISCO; paradata; time-stamps; completion time; survey burden; question adequacy; cognitive capacities of respondents

1 Introduction

The increasing popularity of web-surveys as a new mode of data collection has challenged traditional survey methodology fundamentally. This paper focuses on one feature of web-surveys, namely the measurement of occupations. Occupation is a key variable in socio-economic research, used in studies on labour force composition, social stratification, gender segregation, skill mismatch, and others. Web-surveys are used in commercial market research and in academic fields, like psychology, sociology and health studies, but statistical agencies are cautious to implement web-surveys. Their main focus is on the use of probability-based web-surveys and mixed-mode approaches and little has been done on the implementation of web-surveys (Hoogendoorn & Daalmans 2009). In web-surveys the survey question about occupation particularly is judged risky, as is for example noticed by Statistics Netherlands in an exploration of the use of web-surveys for their Labour Force Survey (Van der Laan & Van Nunspeet 2009). The authors’ worries relate to, among others, breaks in the time series due to changes in the measurement of occupation, and they aim for improvements before using a web-survey for their LFS.

Compared to the variables education and industry, which are also mostly asked in an open response format, the measurement of occupations in web-surveys is the most problematic due to the existence of tens or even hundreds of thousands of job titles. For example, the state of Texas reported over 500,000 distinct job titles in its job evaluation system (Tippins and Hilton 2010). The worldwide, continuous, volunteer WageIndicator web-survey on work and wages has a ten-year experience of measuring occupations, using a search tree and an underlying database of occupations. The web-survey started in the Netherlands in 2001, expanded to other EU member states from 2004 on, included countries outside the EU and in other continents from 2006 on, and today it is operational in almost 60 countries. Each web-survey is in the national language(s) and adapted to peculiarities in the country. Annually hundreds of thousands respondents start to complete the questionnaires. In these web-surveys, respondents self-identify their occupation by means of a 3-tier search tree, allowing them to navigate through the so-called WISCO database with more than 1,700 occupational titles. The search tree has been developed from the need to cope with the occupation variable in a continuous web-survey, generating huge amounts of data. The need became stronger with the expansion to other countries. Search trees are disadvantageous because it poses cognitive demands to the respondent, which might cause dropout. In web-surveys dropout rates are substantial, unknown to other survey modes (e.g. Galesic 2006). This paper details the principles of the database underlying the search tree, the dropout rates during search tree completion and
the time needed to complete the search tree. Note that the paper does not study if the search tree method in web-surveys yields occupation data that are at least as valid or reliable as in other survey modes, because the quality of the search tree itself and its data needs to be investigated first before comparing with other survey modes.

The outline of the paper is as follows. Section 2 reviews the measurement of occupations in the four survey modes Paper-Administered Personal Interviewing (PAPI), Computer-Assisted Personal Interviewing (CAPI), Computer-Assisted Telephone Interviewing (CATI) and Computer-Assisted Web Interviewing (CAWI). Section 3 details the characteristics of the WISCO database and its search tree. Section 4 analyses the dropout rates, the completion time, and the answers to the subsequent open format question. For these analyses, a new dataset has been compiled of the survey data, the time-stamps and the characteristics of the search tree, using the 2010 Q2 WageIndicator web-survey data for the UK, Belgium and the Netherlands. In recent years text string matching has opened new possibilities to measure occupations in web-surveys. Section 5 reviews the requirements for an occupational database underlying this method. Section 6 ends with conclusions.

2 Reviewing the measurement of occupations

2.1 The open and closed response formats in PAPI, CATI, CAPI and CAWI

Many socio-economic surveys, such as Labour Force Surveys and Censuses, include a question “What is your occupation?”, “What kind of work do you do?” or similar, using either an open or a closed response format. Both formats can be used in all four survey modes, but compared to the closed response format the open response format is most often used. Ganzeboom (2010a) advises to use the open format, “because occupations are complicated”. This section summarizes briefly how occupation data is collected in the four modes.

In the open response format question respondents report their job titles as they like, thus this format elicits unstructured data and the data-collector has to undertake field- or office-coding. CAPI and CATI allow for both field- and office-coding, PAPI and CAWI have to rely solely on office-coding. In field-coding, (semi-)automatic indexes can be used to assist interviewers in assigning occupational codes. In office-coding, manual or (semi-)automatic indexes are used. Advanced automatic coding tools allow for classifying huge amounts of unstructured data, and these tools have turned into knowledge databases for field-coding, providing interviewers with a list of potential occupations to choose from (Michiels and Hacking 2004). Automatic indexes, such as CASCOT and its update CASCOT2000 for occupational coding in the UK, are based on previous coding efforts. All indexes are country-specific. As part of the EurOccupations project, a CASCOT multi-language coding tool has become available (Elias, Ellison and Jones 2009). The quality of coding indexes varies across countries. After ILO’s recent update of its international occupational classification ISCO-08, Eurostat has put effort in supporting countries in developing coding indexes for their occupation data collected in Labour Force Surveys and similar.

In the open response format question, part of the respondents tend to report their job title in great detail, as they know it from their employment contract, a job evaluation scheme, or a common understanding in the workplace. Part of these job titles are organization-specific and not understandable beyond the organization. In contrast to detailed job titles, part of the respondents report highly aggregated occupational categories, e.g. ‘clerical worker’ or ‘teacher’, or categories not specific at all, e.g. ‘employee of department X’, ‘senior supervisor’, or ‘dogsbody’. Ganzeboom (2010a) suggest coding crude titles on a 1- or 2-digit ISCO level, using trailing zeroes. In reviewing office-coding, Ganzeboom concludes that this method can lead to substantial percentages of unidentifiable responses and to responses at various levels of aggregation. This is confirmed in the World Values Survey, a predominantly postal survey using office coding for the occupation variable. Its 1999 data for Belgium, the Netherlands, and Great Britain (selection employees and self-employed) reveals that for Belgium the occupation variable is coded only at ISCO88-2 digit and for the Netherlands and Great Britain also at 3- and 4-digit (2-digit: NLD 5%, GBR 8%; 3-digit: NLD 22%, GBR 26%; 4 digit: NLD 72%, GBR 59%). The remaining group is coded missing (BEL 2%, NLD 1%, GBR 6%). This indicates that compared to the missing values, the measurement of
various levels of aggregation is a much larger problem. Note that the data of the unemployed, which are more likely not being able to report an occupational title, have been excluded in these percentages. In the World Values Survey in Belgium their occupations are coded at 2-digit, whereas for the Netherlands and Great Britain the question is considered not applicable for this group.

Taken into account the wide variety of job titles, the occupational dynamics, and the organisation specificity of job titles, it is not surprising that Elias (1997) concludes that office-coding of occupations is an inexact process. Based on a meta-analysis of the results from occupational recoding studies, the author concludes that agreement rates increase with higher levels of aggregation, thus at 1- or 2-digits. At 3-digits, agreement rates in excess of 75 per cent are hard to obtain and 3-digit ISCO comparisons between countries are exposed to the low level of reliability associated with occupational classification. Similarly, Eurostat (2009) notifies that inconsistencies are large for variables that require codification, such as occupations.

In a closed response format question on occupation, a tick list or choice-set offers respondents a choice of occupational titles. This self-identification method can be used in all survey modes and no recoding effort is needed. Yet, in CATI the choice is necessarily limited to 5 - 7 occupational categories, because otherwise the respondents will not memorize all items. These categories are inevitably highly aggregated. PAPI allows for a choice of at most 50 categories, because otherwise the printed questionnaire will exceed a reasonable length. CAPI allows for slightly more categories when using show-cards. CAWI allows for a choice-set of thousands of items, as will be discussed in the next section. A limited choice-set may result in lower data quality, because it is difficult to assure consistency in how respondents fit their own job titles into the highly aggregated categories, introducing aggregation bias (De Vries and Ganzeboom 2008).

2.2 Three response formats in web-surveys (CAWI)

Whereas the PAPI, CATI and CAPI modes basically can employ two response formats, web-surveys facilitate three formats for the survey question about occupations, namely the open response format with office-coding, the closed response format with a search tree, and the open+closed response format with text string matching, the so-called dynamic lists with either auto-completion or suggestions. All three formats will be discussed in this section.

Similar to all other survey modes, web-surveys of course allow for the open response format survey question with office-coding. Due to the absence of an interviewer, however, the results will be comparable to those collected in the PAPI mode and not to those in the CAPI or CATI modes, where interviewers will prevent ambiguous, crude or too detailed responses. In CAWI, the results may be even worse than in PAPI, taken into account the habit of web visitors to key in whatever they like. Hence, when using the open response format in their web-surveys, Statistics Netherlands has to worry even more about their time series, because currently they use predominantly CAPI or CATI for their LFS.

In the closed format survey question, CAWI allows for enlarging the choice-set to thousands of occupational titles compared to the limited choice-sets in PAPI, CATI and CAPI. Respondents can navigate through the choice-set using a search tree with two or three tiers, depending on the number of occupational titles in the choice-set. This so-called multipage filtering is a convenient way to collect data if a variable has too many possible values to be presented on a single page (Funke and Reips 2007). For quite some years, job sites use search trees serving web visitors to identify an occupation. In web-surveys, search trees are advantageous compared to the open response format because aggregation bias and aggregation heterogeneity are prevented and unidentifiable, ambiguous or crude occupational titles are absent. In addition, search trees can easily be applied in multi-country and multi-language surveys, allowing for cross-country comparisons of highly disaggregated occupational data while ensuring comparable survey operations. Yet, search trees are disadvantageous for respondents because they are cognitive demanding and time-consuming, which might lead to substantial dropout rates, a characteristic of web-surveys not present in the other survey modes.

Similar to search engines, the occupation question in web-surveys could use text string matching with multiple suggestions for self-identification, assuming a database underlying the suggestions. A study by Funke and Reips (2007) for a limited set of values showed that these dynamic lists are feasible and that the
response time is lower compared to radio buttons in an experiment offering 48 possible values. Although text string matching may appear to respondents as an open format survey question, it is a closed format question due to the use of a database. To the best of my knowledge, text string matching with self-identification is currently not used for the occupation question in web-surveys. In the Netherlands, text string matching has been, among others, used to categorize millions of job titles in job advertisements, though not into the ISCO classification. The demands to the occupation database underlying text string matching will be elaborated in section 5.

3 The occupation database for search trees

3.1 Introducing the WISCO database

To facilitate respondents’ easy and reliable self-identification of their job title, the 433 units at the 4-digit-ISCO-08 classification are definitely too aggregated. Therefore the WISCO database contains approximately 1,700 unique occupational titles. It consists of a source list of occupational titles (the choice-set), a 3-tier search tree for navigating through the source list, translations into many languages, and 4-digit ISCO-08 codes with follow-up numbers. Translations by national labour market experts have been preferred over translations by professional translators. If a country indicated that two distinct occupational titles in the source list were not considered distinct in that country, one occupation was removed from the country list. In case countries wanted to add occupational titles, these were added to the country list.

For all occupations, search paths have been drafted applying assumptions about respondents’ self-identification behaviour. A main concept is to cluster related occupations in the same 3rd tier of the search tree. In each tier the list of occupations is sorted alphabetically. The search paths facilitate the most efficient searches for occupations with large numbers of jobholders and for occupations with predominantly low-skilled jobholders. Search paths have been designed such that they aim for preventing respondents to ‘upgrade’ their occupations, particular with regard to management occupations. This is detailed in section 3.5.

The WageIndicator web-survey reveals that respondents like to report extensions to their occupational title such as supervision, senior, trainee or others. To satisfy these respondents the web-survey has one follow-up question where they can specify their occupational title using a radio button and one follow-up open format question where they can add additional text information about the occupational title ticked in the search tree. The latter is analysed in section 4.4.

3.2 Efficient search and readability

For respondents the reading time in the search tree should be brief to reduce the dropout risk. Hence, the wording of the occupational titles is kept brief, easy to understand, and unambiguous. The singular has been preferred over the plural and beekeeper over apiarist. No different male and female occupational titles have been used, apart from some countries where this was considered necessary. Synonym titles are not included as these might confuse respondents.

Composite jobs are difficult to measure and to classify. In case of office-coding, composite occupations should be coded according to the highest skill level of any of the composing jobs (Ganzeboom 2010b). Instructions in surveys however usually state that the job title should be ticked where the most time is spent. In web-surveys, the best solution would be to allow respondents to tick two or more occupations. Due to technical constraints the WageIndicator web-survey does not facilitate a second choice, but in the text string matching this method might be feasible.

When coding open response format questions, not all occupational titles can be coded into the classifications’ occupational units and this problem worsens when classifications become outdated due to occupational dynamics in the labour market. The problem is mostly solved by adding residual (“not elsewhere classified”) units. ISCO-08 has 27 residual units. For self-identification these units are
problematic, because respondents will not read the entire choice-set and then conclude that their occupation is not present. However, an occupation database aiming to measure all 4 digit ISCO-08 unit groups has to include the residual units. In the WISCO database the problem has been solved by rephrasing all 27 residual occupation units as “Occupational unit X, all other” and sorting them at the bottom of the appropriate 3rd tier, assuming that respondents have red all occupational titles in that particular 3rd tier before deciding to tick the residual occupation.

3.3 The aggregation level of occupations in the database

The source list of the WISCO database has to optimise between the demand to include as many occupational titles as possible to facilitate reliable self-identification and the demand to be as brief as possible to reduce reading time. To define the aggregation level of occupations, the following definition is employed: “An occupation is a bundle of job titles, clustered in such a way that survey respondents in a valid way will recognize it as at their job title; an occupation identifies a set of tasks distinct from another occupation; an occupation should have at least a not-negligible number of jobholders and it should not have an extremely large share in the labour force” (Tijdens, 2010). The current source list of 1,700 occupational titles turned out to be sufficiently detailed for the vast majority of respondents in a multi-country survey, as can be concluded from the analyses in section 4.4.

Regarding the level of aggregation, a main approach refers to a break-down of occupational titles with large numbers of jobholders into detailed occupational titles, e.g. the ‘clerk’ in tier 2 is broken down into more than ten specific clerk occupations in tier 3. The source list aimed to avoid gender bias, and therefore a number of large female-dominated occupational units have been broken down into detailed occupational titles. The source list is designed such that it leads to valid data without auxiliary variables, and consequently some occupational titles include a reference to industry or firm size. The source list distinguishes handicraft workers from comparable manufacturing workers and retail occupations from manufacturing occupations. For unskilled occupations, broad occupational titles have been preferred. The source list does not distinguish for junior or senior entries of occupational titles, because this would have required an inventory for which occupations in which countries this categorisation is common and for which it is not.

3.4 The skill level of occupations in the database

ISCO-08, as was the case for its predecessors, defines a job as a set of work tasks and duties performed by one person. Jobs with the same set of main tasks and duties are aggregated into the so-called 4-digit occupation units. On the basis of similarity in the tasks and duties performed, the units are grouped into 3- and 2-digit groups, which on the basis of skill level are grouped into 1-digit groups (Greenwood 2004). ISCO distinguishes four skill levels, ranging from unskilled to highly skilled. During the preparation of ISCO-08, the similarity of occupations raised few discussions, but the major discussions concerned the skill levels assumed with the ISCO codes (Elias and Birch 2006).

In the source list, this skill ambiguity is solved by adding national skill requirements to the occupational titles, when known and applicable. For example in the Netherlands the ‘Grafisch vormgever (hbo)’ has been distinguished from the ‘Grafisch vormgever (mbo)’, identifying the graphic designer at medium and at higher vocational training level. For Germany, the ambiguity applied to other occupations, e.g. ‘Archivar/in, Diplom (FH)’, ‘Archivar/in, Diplom (Uni)’, ‘Archivar/in, Fachschule’. Skill requirements have been added to the WISCO country lists when countries indicated a need for it. It turned out to be only relevant in countries where the educational system and the job market are firmly intertwined.

3.5 The corporate hierarchy of occupations in the database

The concepts of careering ad job ladders blur the demarcation lines across occupations, whereas clarity is critical for valid self-identification. For the WISCO database, the corporate hierarchical level of the occupation had to be clarified in the occupational titles, because ISCO has coded these hierarchies at
different skill levels. A stylized six-layer corporate hierarchy has been drafted, assuming that these layers sufficiently would overcome organisation specific settings.

The occupational titles in the source list have been classified as core occupations (OCC), helpers (OCC-1), first line supervisors (OCC+1), heads of departments (OCC+2), managers of institutions, centres, branches and alike (OCC+3), and CEOs, board members and area managers of 50+ organisations (OCC+4). For the heads of department (OCC+2), a stylized horizontal corporate structure of 14 departments has been drafted. For the first-line supervisors (OCC+1), 47 occupational titles have been included, based on a review for which occupations (OCC) first line supervisory positions are likely. For the helpers (OCC-1), almost 20 broad occupational titles have been included. These stylized approaches appeared to be sufficiently detailed to allow self-identification without being forced to draft long lists of specific job titles, as can be concluded from the absence of comments in the open format question in the web-survey with regard to this concept. The search paths have been designed such that the OCC +4 occupations have only one entry through the 1st tier item ‘Management, direction’. Helpers (OCC-1) and first line supervisors (OCC+1) are found in the same tier as the related occupation (OCC).

4 Dropout rates and completion time

4.1 Hypothesis, data and methodology

Given all efforts to design a database of occupations, how adequate is the database for respondents’ self-identification in the WageIndicator web-survey? This section investigates dropout rates, completion time and the responses to the open question following the search tree. In the past decade several studies have attempted to explain dropout and completion time. Basically, the three explanatory clusters to be distinguished are the survey or question burden, the respondents’ interest in the survey topic, and the respondents’ cognitive capacities, particularly reading and computer capacities. The issue of respondents' interest has culminated in attempts for personalisation of surveys. In a study of dropout rates in a web-survey, Galesic (2006) found that the higher respondents’ overall interest in the questions and the lower their overall experienced burden, the lower the dropout risk, thereby asking for interest and burden at the end of each page of the web-survey. Pages that required longer time to complete were more often followed by dropout. The support for the adequacy hypothesis is in line with the findings of Heerwegh and Loosveldt (2006) that personalization has a significant effect on the probability of starting the web-survey and on the probability of reaching and submitting the final web-survey page.

The current study aims to explain dropout rates and completion time of the occupation search tree in the web-survey. It measures survey burden not by using a subjective factor, but from the numbers of characters red in the search tree and from the measure whether respondents have been able to identify their occupations. It measures the interest in the topic not by using a subjective factor, but by assuming that identification with a job title is predominantly applicable to workers in dependent employment and to a lesser extent to self-employed, unemployed, housewives with job on the side, workers in military service and alike. It measures the cognitive capacities from the respondents’ educational attainment. Four research objectives have been studied:

• Are dropout rates during search tree completion explained by the adequacy of the survey question or by the survey burden?
• Is completion time explained by the question adequacy, the survey burden, or the respondents’ cognitive capacities?
• Does completion time of the search tree influences dropout rates after search tree completion?
• How well does the search tree allow respondents to identify their occupation, taken into account the comments posted in the open question following the search tree?

For the analyses, a new dataset has been compiled, derived from the 2010 Q2 WageIndicator web-survey in the United Kingdom, Belgium_Dutch, Belgium_French and the Netherlands. These were the most recent data available at the time of writing. The choice of the three countries was related to the author's language capacities, needed for the third research objective. The new dataset is compiled as follows.
The web-survey data contributes variables about the ticked items in 1st, 2nd and 3rd tier of the search tree, employment status, education level, age, and dropout at the end of part 1 of the questionnaire.

The paradata contributes three time-stamps for completion of the 1st, 2nd and 3rd tier; note that between the first survey page and the page with the 1st tier two other pages are utilized, for which unfortunately no time-stamps are available; note that in case of back-and-forth clicking only the latest time-stamps are recorded.

The occupation database contributes the number of characters including blanks and comma’s in the most efficient search path for each ticked item in the 2nd tier and in the 3rd tier, assuming no further reading once respondents have identified their occupations.

The text data of the web-survey contributes the text that respondents have keyed in in the open question following the search tree. The author has coded these responses. From this data a variable called ‘wrong match’ is derived, indicating that respondents have keyed in another occupation in the follow-up open survey question than ticked in the search tree (see section 4.4 for more details).

4.2 Dropout rates during search tree completion

Page 1 of the web-survey asks a question about employment status and in 2010q2 in the three countries 24,811 respondents have provided a valid answer. Pages 2a and 2b ask a few questions to respondents with respectively without a job. Dropout on these pages is 4.9%. Page 3 asks in which region the respondent lives, using a drop down format question. Dropout on this page is 2.4%. Pages 4, 5 and 6 consist of the three tiers of the occupation search tree. Completing the search tree is compulsory. The number of respondents entering page 4 is 22,990. For research objective 1 the dependent variables are the dropout rates at each tier of the search tree, as shown in Table 1. The Table reveals that one in five respondents drop out during search tree completion and that more than half of the dropout takes place on the 1st tier of the search tree. The percentages vary slightly across countries.

<table>
<thead>
<tr>
<th>Country/Language Combination</th>
<th>UK</th>
<th>Belgium (fr)</th>
<th>Belgium (nl)</th>
<th>Netherlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>N when entering search tree</td>
<td>1,611</td>
<td>1,515</td>
<td>2,278</td>
<td>17,586</td>
</tr>
<tr>
<td>did not complete 1st tier</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>did not complete 2nd tier</td>
<td>10.3%</td>
<td>11.9%</td>
<td>11.2%</td>
<td>15.0%</td>
</tr>
<tr>
<td>did not complete 3rd tier</td>
<td>4.4%</td>
<td>2.7%</td>
<td>3.2%</td>
<td>2.4%</td>
</tr>
<tr>
<td>total dropout in search tree</td>
<td>21.4%</td>
<td>18.1%</td>
<td>18.9%</td>
<td>21.9%</td>
</tr>
</tbody>
</table>

Source: WageIndicator survey, Belgium, UK, Netherlands, 2010q2 (N= 22,990 with valid observations at the search tree entry)

 Dropout is hypothesized to be influenced by the adequacy of the survey question and the survey burden. Question adequacy is defined as the occupation survey question being most adequate for employees, lesser to self-employed and not at all to other, marginal groups in the labour market. Survey burden is defined as the number of characters in the most efficient search path, here the search path until the dropout page. A logistic regression model is used to investigate dropout probabilities. Controls for country are not included because they did not reveal significant results. Note that the impact of the cognitive capacities and the ‘wrong match’ cannot be studied here because this information is not available for the respondents that dropped out.

Panel 1 in Table 2 reveals support for the question adequacy hypothesis. Being an employee and being self-employed lowers the odds ratio of the dropout probability substantially. For the 1st tier the length of the relevant search path is by definition not available and thus the burden hypothesis is not investigated. Panel 2 reveals neither support for the question adequacy hypothesis nor for the burden hypothesis. Panel 3 shows no support for the question adequacy hypothesis but it does for the burden hypothesis. Both the number of characters in the 1st and the 2nd tier increase the odds ratio of the dropout probability.
In summary, previous findings that the dropout rate is largest at the beginning of the questionnaire (Galesic 2006) are not confirmed here. In the WageIndicator web-survey the dropout at page 4 is substantially higher than at pages 2 and 3. The question adequacy hypothesis is partly supported, because it contributes to the dropout explanation in tier 1, but not to that in tier 2 and tier 3. This is understandable, because respondents will identify the adequacy of the question for their particular situation at the 1st tier of the search tree and not at the 2nd or 3rd tier. Galesic’s burden findings are partly confirmed. In the WageIndicator search tree, the survey burden contributes to the dropout explanation in tier 3, but not in tier 2.

4.3 Time needed to complete the search tree

In research objective 2 the completion time in each tier of the search tree is the dependent variable. For each completed tier a time-stamp is available, but unfortunately no time-stamp is available for the last question before starting the search tree. Therefore no completion time is computed for the 1st tier. The completion time of the 2nd tier and the 3rd tier is computed in seconds. A few observations with completion times of less than 0 seconds or more than 500 seconds per tier have been considered out of range. Table 3 shows that the median completion time for the 2nd tier is 11 seconds and for the 3rd tier it is 14 seconds. By definition, no completion time is computed for the cases that dropped out in the tier at stake. Total completion time of the 3-tier search tree might be between 30 and 60 seconds.

Table 3 Descriptive statistics of the completion time in seconds of completing tier 2 and tier 3 in the search tree

<table>
<thead>
<tr>
<th>Name</th>
<th>N</th>
<th>% of total obs</th>
<th>Min.</th>
<th>Max.</th>
<th>Median</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completion time tier 2</td>
<td>18,805</td>
<td>99.60%</td>
<td>1</td>
<td>492</td>
<td>11</td>
<td>16.59</td>
<td>20.2</td>
</tr>
<tr>
<td>Completion time tier 3</td>
<td>17,680</td>
<td>99.40%</td>
<td>1</td>
<td>495</td>
<td>14</td>
<td>19.51</td>
<td>23.83</td>
</tr>
</tbody>
</table>

Source: WageIndicator survey, Belgium, UK, Netherlands, 2010q2

Question adequacy, survey burden and respondents’ cognitive capacities are the explanatory variables. Question adequacy has been defined similar as in the previous section. Survey burden includes the number of characters red in the most efficient search path and an indication for ‘wrong match’. Respondents’ cognitive characteristics have been defined as their educational level and their age. OLS regressions have been applied to investigate completion time. Due to dropout, the number of observations is for education and age lower than for the variables related to the occupation search tree. Therefore, model 1 estimates completion time without education and age and model 2 does so with education and age.

Table 4 panel 1 reveals that the question adequacy hypothesis is partly supported because employees indeed need less time to complete tier 2 than the self-employed and the marginal groups in the labour market, but the effect slides into insignificance in model 2. In both models the survey burden hypothesis is supported, showing that the more characters in tier 2, the longer the completion time. Yet, respondents who have read more characters in tier 1 spend less time in completing tier 2. Respondents with a wrong match spend substantial more time in completing tier 2. The cognitive characteristics hypothesis is

---

**Table 2** Effect of employment status and number of characters in the search tree on the probability of dropping out (0=no dropout, 1=dropout)

<table>
<thead>
<tr>
<th>Odds ratio tier 1 dropout probability</th>
<th>Odds ratio tier 2 dropout probability</th>
<th>Odds ratio tier 3 dropout probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee*</td>
<td>0.152 ***</td>
<td>0.875</td>
</tr>
<tr>
<td>Self-employed*</td>
<td>0.185 ***</td>
<td>0.729</td>
</tr>
<tr>
<td># characters in tier1 (41-839)</td>
<td>1.000</td>
<td>1.001 ***</td>
</tr>
<tr>
<td># characters in tier2 (4-307)</td>
<td></td>
<td>1.002 ***</td>
</tr>
<tr>
<td>Constant</td>
<td>0.481 ***</td>
<td>0.031 ***</td>
</tr>
<tr>
<td>-2 Log likelihood</td>
<td>16455.60</td>
<td>5420.57</td>
</tr>
<tr>
<td>N</td>
<td>22,992</td>
<td>19,526</td>
</tr>
</tbody>
</table>

Source: WageIndicator survey, Belgium, UK, Netherlands, 2010q2

* Reference categories: marginal groups other than employee or self-employed
*** p<0.001, ** p<0.005; * p<0.010
supported for education and age. The high educated need less completion time compared to the middle educated and the older respondents need more completion time.

Table 4 panel 2 reveals that the question adequacy hypothesis is not supported. The number of characters in the 1st and the 2nd tier are not relevant for the completion time in the 3rd tier, except in model 2 for the 1st tier. In both models the survey burden hypothesis is supported, because more characters in tier 3 cause longer completion time. Respondents with a wrong match spend substantial more time in completing tier 3. The cognitive characteristics hypothesis is supported for education and age. Completion time is longer for the low educated compared to the middle educated and the older respondents need more completion time.

Finally, hypothesis 3 addresses if completion time in the search tree influences dropout at the end of part 1 of the questionnaire? Table 5 holds the results of a logistic regression analysis, revealing that indeed the time-consuming search tree is forgotten at the end. The table shows support for the question adequacy hypothesis, although here this would better be phrased as the survey adequacy hypothesis. For employees the log odds of dropping out is substantial lower compared to the marginal groups in the labour market. A reason might be that the survey questions mainly address issues on work and wages. The model also supports the hypothesis regarding the cognitive demands of the survey. The log odds of the dropout probability increases for the low educated compared to the middle educated, whereas it is substantial lower for the higher educated compared to the middle educated.
Table 4  Effect of employment status, number of characters in the search tree, match, education and age on the seconds needed to complete tier 2 and tier 3 in the occupation search tree (unstandardized coefficients and standard errors of OLS regressions)

<table>
<thead>
<tr>
<th>Completion time tier 2</th>
<th>Completion time tier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td>(Constant)</td>
<td>14.955</td>
</tr>
<tr>
<td>Employee*</td>
<td>1.223 ***</td>
</tr>
<tr>
<td>Self-employed*</td>
<td>0.789</td>
</tr>
<tr>
<td># characters in tier 1 (41-839)</td>
<td>-0.005 ***</td>
</tr>
<tr>
<td># characters in tier 2 (4-307)</td>
<td>0.027 ***</td>
</tr>
<tr>
<td># characters in tier 3 (6-2456)</td>
<td>0.020 ***</td>
</tr>
<tr>
<td>Wrong match according to text</td>
<td>5.522 ***</td>
</tr>
<tr>
<td>Education low*</td>
<td>-0.100</td>
</tr>
<tr>
<td>Education high*</td>
<td>-1.997 ***</td>
</tr>
<tr>
<td>Age (10-80)</td>
<td>0.081</td>
</tr>
<tr>
<td>Adj Rsquare</td>
<td>0.016</td>
</tr>
<tr>
<td>N</td>
<td>18,742</td>
</tr>
</tbody>
</table>

Source: WageIndicator survey, Belgium, UK, Netherlands, 2010q2

* Reference categories: marginal groups other than employee or self-employed; education middle

*** p<0.001, ** p<0.005; * p<0.010
### Table 5: Effect of employment status, completion time, match, education and age on the probability of dropping out at the end of part 1 of the questionnaire (0=no dropout, 1=dropout)

<table>
<thead>
<tr>
<th></th>
<th>Odds ratio part 1 drop out probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee*</td>
<td>.678***</td>
</tr>
<tr>
<td>Self-employed*</td>
<td>1.111</td>
</tr>
<tr>
<td>Completion time tier 2</td>
<td>.999</td>
</tr>
<tr>
<td>Completion time tier 3</td>
<td>1.000</td>
</tr>
<tr>
<td>Wrong match according to text</td>
<td>.778</td>
</tr>
<tr>
<td>Education low*</td>
<td>1.358***</td>
</tr>
<tr>
<td>Education high*</td>
<td>.864**</td>
</tr>
<tr>
<td>Age</td>
<td>1.001</td>
</tr>
<tr>
<td>Constant</td>
<td>.989</td>
</tr>
</tbody>
</table>

-2 Log likelihood: 14999.160
Chi-square (df 8) ***: 198.227
N: 11097

Source: WageIndicator survey, Belgium, UK, Netherlands, 2010q2
* Reference categories: marginal groups other than employee or self-employed; education middle
*** p<0.001, ** p<0.005; * p<0.010

### 4.4 Analyses of the response to the open response format question

Research objective 3 focuses on how well the search tree allows respondents to identify their occupation, taken into account the comments posted in the open response format question following the search tree asking if they wanted to add anything to the search tree. In total, 4,020 respondents have keyed in relevant text in the open question (22.6% of the 17,782 who completed the 3rd tier of the search tree). Relevant text is defined as text which includes at least two letters and is not a ‘no’ response to the question. Particularly in Belgium, this percentage is relatively high (29.6% for fr_BE and 50.4% for nl_BE), whereas it is almost equal for Netherlands and United Kingdom (18.9% respectively 16.7%). This may indicate that even though Flanders and the Netherlands have the same language, the Flanders job titles might be different from those in the Netherlands, where the national list has been drafted. The ticked occupational title and the answers in the text box have been classified in seven categories (Table 6).

### Table 6: The categories used to classify the responses to the open response follow-up question on occupation in comparison to the ticked occupation and their frequencies.

<table>
<thead>
<tr>
<th>Match category</th>
<th>Explanation</th>
<th>% of valid text boxes</th>
<th>% of ticked occ. titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERFECT</td>
<td>Text and ticked occupational title were similar</td>
<td>3.6%</td>
<td>0.81%</td>
</tr>
<tr>
<td>ADDITIONAL</td>
<td>Text provides additional information to ticked occupational title</td>
<td>69.7%</td>
<td>15.8%</td>
</tr>
<tr>
<td>50% MATCH</td>
<td>Text indicates that ticked occupational title is not wrong, but the search tree has better alternatives</td>
<td>13.5%</td>
<td>3.1%</td>
</tr>
<tr>
<td>WRONG</td>
<td>Text indicates that ticked occupational title is wrong</td>
<td>7.5%</td>
<td>1.7%</td>
</tr>
<tr>
<td>IRRELEVANT</td>
<td>Text is irrelevant for ticked occupational title</td>
<td>4.8%</td>
<td>1.1%</td>
</tr>
<tr>
<td>GENERAL</td>
<td>Text includes an aggregated occupational title compared to ticked occupational title</td>
<td>0.9%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100%</td>
<td>22.6%</td>
</tr>
</tbody>
</table>

Source: WageIndicator survey, Belgium, UK, Netherlands, 2010q2

The category ADDITIONAL includes either extended task descriptions or refers to composite jobs. An example is: ‘I am a secretary with HR tasks’. An example of 50% MATCH is when the ticked title is ‘civil servant in a municipality’ and the text box states that the respondent has a clerical job. An example of IRRELEVANT is ‘I like my job but not my boss’. The category GENERAL is particularly used in Belgium, where respondents refer to the distinction between blue and white collar, which is relevant in the country. Table 6 shows that by far the most texts fall in the category ADDITIONAL, challenging the
occupational boundaries being not as distinct as the search tree and the occupational classification assume. This problem can basically be solved by facilitating a second choice in the search tree.

The category WRONG reveals that the text includes another occupation than the one ticked in the search tree, thus a wrong match between the search tree data and the text question. In total 1.7% of all ticked occupations are obviously wrong, requiring an occupational recoding of the variable derived from the search tree. The WRONG percentages are not equally distributed over the occupational titles in the search tree. The four titles with the most frequent wrong answers (only occupations with at least 10 observations) are ‘Craft or related worker, all other’, ‘Paramedical practitioner, all other’, ‘Process controller, all other’, and ‘Sales representative’. Similar to the residual occupations in office coding, in search trees in web-surveys the category ‘all other’ fills easily. For these four occupations, the search paths in the occupation database need revision.

In summary, given the percentage of 1.7% for wrong ticked occupations, the search tree is likely to yield reliable data. The problem of composite occupations needs a solution. Finally, lists of occupational titles are country-specific and even when these countries have the same language, they need to be checked, as is shown for Flanders, using the Dutch list of occupational titles.

5 Database requirements for text string matching

Summarizing the findings so far, the search tree does not appear the most optimal response format for the occupation question in web-surveys, because of the not-negligible dropout rates and the time needed to complete the search tree. Though, a search tree is probably as good as the open format question with office-coding. The latter is anyhow problematic for continuous, multi-country web-surveys, particularly in case of the absence of an associated national institute undertaking the recoding. The text string matching method is a promising alternative, particularly when aiming for high quality text string matching for multi-country and multi-language use in web-surveys.

Text string matching is neither similar to the current open format questions with (semi)-automatic office indexes nor to the search tree method. The three methods have in common that they classify occupations, either into ISCO-08 or in another classification. In contrast to a search tree database, a coding index and a text string matching database have to cope with unstructured text. A coding index classifies each response into one occupational title in the classification and an advanced coding index provides match percentages for several occupations. A text string matching database should provide a list of the 5 to 10 best matching occupational titles to the respondent’s text, because these 5 to 10 titles should be presented on the screen to allow the respondent to select the best match. Whereas a search tree database includes at most a few thousand occupational titles, a coding index might count tens of thousands of job titles, and a text string matching database will be in between. Similar to a coding index, a text string matching database should provide matches for disaggregated job titles including rare titles, whereas search trees do not include rare titles, because most respondents are capable of aggregating their rare job title into an occupational title present in the search tree, as can be concluded from the open response question, discussed in the previous section. A text string matching database will have to extend the current advanced coding indexes with algorithms for the selection of the 5 to 10 best matching occupational titles.

A first group of algorithms is needed for cleaning the text keyed in by the respondent. For cleaning text in Latin scripts, algorithms are needed for coping with standard typing errors, blanks, weird signs and a like, for reversing plural titles into singular titles and female titles into male titles, for deleting numbers added to job titles such as ‘Clerk 1’ or ‘2nd cook’, and for deleting additions such as senior, junior, supervisor and alike. For the latter, for many countries such a list is available, because the WageIndicator web-survey has a survey question on this topic.

A second group of algorithms needs to communicate with respondents using ambiguous, crude or highly aggregated occupational titles, asking them to specify their job title in greater detail. This applies to ambiguous words such as ‘employee’, ‘supervisor’, ‘junior’, ‘senior’, to core parts of job titles, e.g. ‘maker’, ‘operator’ and to unidentified phrases such as ‘dogsbody’ or ‘I don’t like my boss’. It also applies to crude occupational titles, e.g. ‘engineer’, ‘clerk’, ‘nurse’, ‘teacher’, because these titles will generate too
many matches. Hence, the database should include a list of the most common ambiguous, crude or highly aggregated occupational titles, including a reply script for communication with the respondent.

A third group of algorithms needs to generate the 5-10 best matching occupational titles for each text string. It needs algorithms that for each best matching occupational title drafts a list of the related occupation(s) in the corporate hierarchy, the comparable occupation(s) in neighbouring skill levels and the neighbouring occupations. These algorithms are important for a valid self-identification of professionals versus semi-professionals, proper skill levels, and occupations with blurred boundaries. Of course these algorithms may lead to more than ten occupational titles. This finally requires an algorithm that selects the matches with the highest match scores, preferably based on frequencies of previous text string matches.

6 Conclusions

Occupation is a key variable in socio-economic research. Web-surveys can employ three methods for the occupation question, namely a closed response format using a search tree, an open response format with office coding, and a combined open+closed format using text string matching. Web-surveys are disadvantageous because unidentifiable or aggregated responses can occur due to the absence of an interviewer. To facilitate respondent’s self-identification, the multi-country WageIndicator web-survey employs a 3-tier search tree with 1,700 occupations. The empirical findings in this paper show that the search tree apparently is not the most optimal response format for the occupation question in web-surveys. Dropout rates are approximately 20% in the 3-tier search tree. Respondents spend on average between 30 to 60 seconds on search tree completion. The percentage of wrong ticked occupations is 1.7%, indicating that the search tree is likely to yield reliable data. The follow-up open response question shows that the problem of composite jobs needs to be tackled, which could be solved by allowing more than one response in the search tree. The analysis shows that lists of occupational titles are country-specific and even when these countries have the same language, they need to be checked for country-specificity, as is shown in the large number of comments of respondents from Flanders where the Dutch list of occupational titles is used.

Using the 2010q2 WageIndicator data for UK, Belgium and Netherlands (24,811 observations), this paper investigated search tree dropout rates, completion time and respondents’ use of an open question following the search tree. Hypotheses relate to question adequacy (do respondents identify a job title), to survey burden (number of characters in respondent’s search path and respondent could find his occupation in the search tree), and to respondents’ cognitive capacities.

The question adequacy hypothesis explains the dropout rate in the 1st tier of the search tree, but not in the 2nd and 3rd tier. This is understandable, because most respondents will identify the adequacy of the question for their particular situation in the 1st tier. The hypothesis hardly explains completion time in the 2nd and 3rd tier.

The burden hypothesis does not explain dropout rate in 2nd tier, but it does so in the 3rd tier, because here dropout increases with the number of characters in the 1st and 2nd tier. The hypothesis does also explain the completion time in both tiers. The more characters in tier 1, 2 and 3, the longer the completion time. Respondents with a wrong match after search tree completion have spend substantial more time in completing tier 2 and tier 3.

The cognitive characteristics hypothesis is supported for education and age, as in the 2nd tier the high educated need less completion time and in the 3rd tier the low educated need more time. In both tiers the older respondents need more time to complete the tier.

A last study investigated if total completion time in the search tree influenced dropout at the end of part 1 of the questionnaire. The results show that in the end the time-consuming search tree is forgotten. The overall conclusion is that in case of multi-country continuous web-surveys the use of a search tree for the occupation question has some serious drawbacks, but that the alternative of an open survey question with office-coding has also serious drawbacks. Text string matching is assumed to overcome these drawbacks, but needs further development efforts, preferably on a cross-country scale.
References


