Ca' Foscari University of Venice

Department of Economics

Working Paper

Michele Belloni,
Agar Brugiavini,
Elena Meschi, and
Kea Tijdens

Measurement error in occupational coding: an analysis on SHARE data

ISSN: 1827-3580
No. 24/WP/2014

Working Papers
Department of Economics
Ca' Foscari University of Venice
No. 24/WP/2014
ISSN 1827-3580

Measurement error in occupational coding: an analysis on SHARE data

Michele Belloni
Department of Economics, University of Venice at Ca' Foscari

Agar Brugiavini
Department of Economics, University of Venice at Ca' Foscari

Elena Meschi
Department of Economics, University of Venice at Ca' Foscari

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

Abstract: This article studies the potential measurement errors when coding occupational data. The quality of occupational data is important but often neglected. We recoded open-ended questions on occupation for last and current job in the Dutch SHARE data, using the CASCOT ex-post coding software. The disagreement rate, defined as the percentage of observations coded differently in SHARE and CASCOT, is high even when compared at ISCO 1-digit level (33.7% for last job and 40% for current job). This finding is striking, considering our conservative approach to exclude vague and incomplete answers. The level of miscoding should thus be considered as a lower bound of the "true" miscoding. This highlights the complexity of occupational coding and suggests that measurement error due to miscoding should be taken into account when making statistical analysis or writing econometric models. We tested whether the measurement error is random or correlated to individual or job-related characteristics, and we found that the measurement error is indeed more evident in ISCO-88 groups 1 and 3 and is more pronounced for higher educated individuals and males. These groups may be sorted in occupations that are intrinsically more difficult to be classified, or education and gender may affect the way people describe their jobs.

Keywords: occupation, ISCO, disagreement rate, coding software, gender, education

JEL Codes: C81, C88, J01, J21 J82

Address for correspondence:
Elena Meschi
Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta S.Giobbe
30121 Venezia - Italy
Phone: (++39) 041 2349153
Fax: (++39) 041 2349176
e-mail: elena.meschi@unive.it

The Working Paper Series
is available only on line
(http://www.unive.it/nqcontent.cfm?a_id=86302)
For editorial correspondence, please contact:
wp.dse@unive.it

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta San Giobbe
30121 Venice Italy
Fax: ++39 041 2349210

Measurement error in occupational coding: an analysis on SHARE data[1]

Michele Belloni
Department of Economics,
University Ca' Foscari of Venice

Agar Brugiavini
Department of Economics
University Ca' Foscari of Venice

Elena Meschi
Department of Economics,
University Ca' Foscari of Venice

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

Abstract

This article studies the potential measurement errors when coding occupational data. The quality of occupational data is important but often neglected. We recoded open-ended questions on occupation for last and current job in the Dutch SHARE data, using the CASCOT ex-post coding software. The disagreement rate, defined as the percentage of observations coded differently in SHARE and CASCOT, is high even when compared at ISCO 1-digit level (33.7% for last job and 40% for current job). This finding is striking, considering our conservative approach to exclude vague and incomplete answers. The level of miscoding should thus be considered as a lower bound of the "true" miscoding. This highlights the complexity of occupational coding and suggests that measurement error due to miscoding should be taken into account when making statistical analysis or writing econometric models. We tested whether the measurement error is random or correlated to individual or job-related characteristics, and we found that the measurement error is indeed more evident in ISCO-88 groups 1 and 3 and is more pronounced for higher educated individuals and males. These groups may be sorted in occupations that are intrinsically more difficult to be classified, or education and gender may affect the way people describe their jobs.

Keywords: disagreement rate; ISCO; coding software; gender; education

[1] This paper builds on research work done for the DASISH project (Data Service Infrastructure for Social Sciences and Humanities, funded by EU-FP7, Contract no. 283646) and the SHARE project (Survey of Health, Ageing and Retirement in Europe). The authors are grateful to researchers at CentERdata in the Netherlands for their coding efforts of Dutch job titles. The authors would like to thank Eric Balster, Peter Elias, Eric Harrison, Maurice Martens, Sue Westerman, and all participants of "CASCOT: Occupational Coding in Multi-national Surveys" Workshop in Venice (10-11 April 2014) for helpful suggestions on earlier draft of this work. This paper benefited from a research visit of Michele Belloni to the University of Amsterdam, made possible through a visiting grant of the InGRID project (Inclusive Growth Infrastructure Diffusion Research Infrastructure, funded by EU-FP7, Contract no. 312691). The last author would like to acknowledge the contribution of WEBDATANET [COST Action IS1004].
This paper uses data from SHARE wave 1 release 2.5.0, as of August 23th 2011. The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5-CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions)."

1. Introduction

Knowledge of individuals' occupation is an important information for many studies in social sciences. For instance in economics, sociology, and other disciplines occupation is often considered, either itself or as part of an index, as a proxy for socioeconomic status. In labour economics, occupation is a key variable in a wide strand of studies, such as the "task approach" to labour markets and job polarization (e.g. Autor 2013; Autor et al. 2006; Goos and Manning, 2007), the definition of skill mismatch and over-education (for extensive overview of this literature e.g. Hartog 2000; Leuven and Oosterbeek 2011), and analysis of the effect of occupation on health status (e.g. Fletcher et al. 2011; Ravesteijn et al. 2013).

In this literature, the quality of occupational data is hardly discussed, despite the fact that measuring occupation in social surveys is a rather complex issue. Handbooks detail how to ask for occupation in Labour Force Surveys and Censuses, among others by international organizations such as the International Labour Organization (ILO) (e.g. ILO, 2010). However, empirical research on best practices and on miscoding is little. The difficulty to provide researchers with an accurate measure of occupation firstly regards the choice of the question(s) to include in the questionnaire and the related training to interviewers and then relates to the conversion of job tiles, that are often recorded as open text field into occupational codes.

The statistical agencies of 150 countries associated in the ILO have adopted the International Standard Classification of Occupations (ISCO) to harmonize the measurement of occupations. The first classification dates back to 1958, with updates in 1968, 1988 and recently in 2008. The Commission of the European Communities (2009) has adopted ISCO-08 as its occupational classification, and the European statistical agency Eurostat has put effort in supporting European countries in developing coding indexes for their occupation data collected in Labour Force Surveys and similar surveys. In 2012 almost half of the 150 countries used ISCO with the other half either not classifying occupations or maintaining an own classification (UN 2014).

The ILO provides a classification and task descriptions for all 4-digit occupational units in ISCO (ILO, 1014). The task descriptions provide also a coding index, but only in English. Therefore, coding occupations becomes particularly challenging in international surveys, such as the *Survey of Health, Ageing and Retirement in Europe* (SHARE) and the *European Social Survey* (ESS), where the occupational codes should be fully comparable across countries, because it is sometimes problematic for countries to map their specific occupations and job titles into the international ISCO categories.

Researchers are often not aware of the complex preparatory work behind occupational coding. They consider the published variable 'occupation' as free of error. In this article, we will first point out that this might not be the case. In addition, we will test whether such a measurement error in occupation is random or is instead correlated to some specific individual or job-related characteristics. We suggest to take this potential measurement error in occupation into consideration when making statistical analysis or writing econometric models.

To reach these aims, we conduct the following empirical analysis. We recode open-ended questions on occupation for the Dutch sample of SHARE data using a well-known software for ex-post coding called CASCOT. We then compare SHARE originally published with recoded occupational variables. Finally, we analyse which individual characteristics (such as gender, education, or industry) are associated to the probability of different coding. The article proceeds as follows: Section 2 discusses the alternative methods used to collect and code information on individuals' occupations and describes the main features of CASCOT. In section 3, we describe our empirical exercise and present the data and the methodology adopted. The results of our analysis are presented and discussed in section 4. Finally, section 5 concludes and suggests some directions for further research.

2. Coding occupations in survey data: alternative methods

Most of occupational information in survey data is obtained from direct questions addressed to respondents. The question about occupation is usually asked as an open text field (e.g.: "What occupation did you perform in your principal job during the week of … to … ?") (see for an overview of survey questions Tijdens 2014b). Occupation can also be asked using a tick list, where respondents have to self-classify in a list of occupational titles. Depending on the survey mode, this list consists of a limited set of necessarily broad occupational groups in mail surveys or lists of thousands of items in web surveys. The main advantage with self-classification (or self-coding) is that surveys do not need a costly and time-demanding coding process. There are, however, many shortcomings with self-coding. 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, thereby introducing aggregation bias (De Vries and Ganzeboom 2008). Both the validity (correct categorization) and the reliability (same categorization made by different interviewers of equivalent responses) of pre-coded occupational categories have been shown to be very poor. An extensive look-up table with a search tree leads to drop-out in web-surveys, but this problem may be tackled in case of text string matching (Tijdens 2014a). Promising attempts to code job titles *during* CAPI interviews are being made, using a look-up table or coding index. SHARE is currently testing a semantic text string matching algorithm developed by CentERdata (http://www.centerdata.nl/) for possible use in its future data collection.

Most surveys however still use an open-ended question with occupational coding (for question design see Jackle 2008). In its handbook for the measurement of the active population in censuses the ILO provides detailed instructions for the use of an open-ended questions and the ILO does not consider self-coding as an alternative (ILO, 2010). Open-ended questions allow classifying occupations at a detailed level of disaggregation, but the text fields require recoding afterwards ('office coding'). The classification of occupational information is in fact achieved through a coding process that converts the reported job titles into a set of codes and that can be done manually or semi-automatically, using a computerised coding system ('computer-assisted coding') or by a combination of both. Manual coding requires a lot of training for coders and coders supervisors (see Hoffmann, Elias, Embury and Thomas, 1995; Ganzeboom, 2008). Semi-automatic coding tools are becoming more and more reliable instruments using semantic matching with previously coded occupations. Recently, machine learning algorithms appear to be a promising development, requiring a substantial amount of manually coded occupations to be used as training data for the automatic classification (Bethmann et al 2014; Cheeseman Day 2014).

CASCOT is a software tool for coding text automatically or manually (http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/) developed at the Institute for Employment Research (IER) in 1993 and since then continuously updated and used by over 100 organisations in the UK and abroad. The software developed at IER is able to code job titles into UK various editions of Standard Occupational Classification (SOC) and International Standard Classification of Occupations (ISCO)[2]. CASCOT software is coupled with an editor which allows users to modify internal coding rules and allows the software to use alternative occupational classification structures. A high quality coding requires high quality job descriptions. The recorded text should ideally contain

[2] An international version of CASCOT, which will allow to code occupations in many languages and multi-national surveys, is under development within the EU financed project DASISH (see http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/internat/).

sufficient information to distinguish it from alternative text descriptions which may be coded to other categories within the classification, but it should not contain superfluous words. This ideal will not always be met but CASCOT has been designed to perform a complicated analysis of the words in the text, comparing them to the words in the classification, in order to provide a list of recommendations. If the input text is not sufficiently distinctive, it may not be the topmost recommendation that is the correct code. When CASCOT assigns a code to a piece of text, it also calculates a score from 1 to 100 which represents the degree of certainty that the given code is the correct one. When CASCOT encounters a word or phrase that is descriptive of occupation but lacks sufficient information to distinguish it from other categories (i.e. without any further qualifying terms) CASCOT will attempt to suggest a code but the score is limited to below 40 to indicate the uncertainty associated with the suggestion (for example cases like 'Teacher' or 'Engineer'). The performance of CASCOT has been compared to a selection of high quality manually coded data. The overall results show that 80% of records receive a score greater than 40 and of these 80% are matched to manually coded data. When using CASCOT one can expect this level of performance with similar data, but the performance depends on the quality of input data. For more information about the software, see Elias et al. (1992) and Jones and Elias (2004).

The user may run CASCOT in three different modes: fully automatic, semi-automatic, and manual or one-by-one. The fully automatic mode does not require any human intervention once a list of job descriptions is provided to the software: a series of corresponding codes plus the associated scores is produced; if the software considers the quality of a given job description too low to be impossible for it to attribute any reasonable code, it provides “no conclusion” for that specific text. The semi-automatic mode works by setting a minimum score: in all cases in which CASCOT attributes a score greater than the minimum value, it codes the text automatically; otherwise it asks for human intervention. The operator, in these cases, is asked to choose manually between a list of recommendations. In manual mode, for each job description, CASCOT provides a list of recommended codes with corresponding scores and leaves the final choice of the best code to the operator. Although time consuming, this mode ensures the maximum level of control on the output. Obviously, the operator tends to choose the topmost recommendation when the score is high and concentrates on the cases which show lower scores.

A Dutch version of CASCOT has been developed at Statistics Netherlands (CBS) building upon its English version. Since 2012, this software (CASCOT-NL henceforth) has been used in the

Netherlands to code job titles in the most relevant social surveys including the Dutch Labor Force Survey. CASCOT-NL is suitable for implementation in CAPI, CATI and CAWI-modes.

In this study, we use a version of CASCOT-NL which CBS used from 01-04-2012 until 01-04-2013 to code job descriptions into 4-digits ISCO-08 in its Labour Force Survey. A noticeable difference between CASCOT-UK and CASCOT-NL (so called "classification file ISCO v1.1") is that the latter includes a special category for vague responses, called "99..". This is because - once tagged in this way - these especially problematic answers go through subsequent coding steps. These steps exploit information from additional variables such as sector of work, individuals' educational attainments and tasks and duties involved in the job; finally, the most difficult cases are manually coded by a team of experts. See CBS (2012) and Westerman (2014) for further details on CBS coding procedures.

3. Data and empirical strategy

Our analysis is based on SHARE data. SHARE is a cross-national longitudinal survey on health, socio-economic status and social and family networks representative of the population aged 50 and over. Four waves of SHARE are currently available. We focus on the first wave of the data (collected in 2004-2005), because this is the only one in which information on occupation was gathered through an open-ended question. In particular, in SHARE wave 1 respondents were asked the following question: "What is your [main/last] job called? Please give the exact name or title". This question was asked to both employed/self-employed and retired/unemployed individuals (the latter conditional on having worked earlier in life).[3]

SHARE country teams manually coded the text strings on respondents' job titles into ISCO-88 (COM) - the International Standard Classification of Occupations in place at that time. Each country team hired and trained coders independently. Coders were asked to follow a protocol providing them with guidelines on how to code "critical" jobs (e.g. managers in agriculture or teachers). These guidelines were partly common to all countries, and partly language-specific. SHARE coders made

[3] SHARE also collects information on respondents' second job, parents' job and former partner's job. Parents' jobs are intrinsically more difficult to code than respondents' jobs because the former may have been excluded from recent job classifications. There are very few observations for respondents' second job and former partner's job. Thus, we exclude these additional variables from our analysis.

also use of ancillary information on training and qualifications needed for the job (this last information was not included in the public release of the data) and on the industry the respondent was working in, based on the question "What kind of business, industry or services do you work in (that is, what do they make or do at the place where you work)?". From one side, SHARE coders were asked to code job descriptions at the maximum possible level of detail, i.e. at 4-digit or 'unit group' ISCO-88 level. On the other side, they were suggested to code vague responses by means of trailing zeros: this means that in case they were unsure if a given job description could have been attributable to a given unit group, they should have attribute it to either a minor (i.e. 3-digits), sub-major (2-digits) or major (1-digit) group. Two variables - one for *current main job (ep016_*) and one for *last job* (*ep052_*) - reporting generated ISCO-88 codes were finally published (for further details, see MEA, 2013, p. 29).

The first wave of SHARE covers 11 European countries, plus Israel. Our recoding exercise exploits only the Dutch sample of this wave, because CASCOT is currently available in two languages - English and Dutch - and the English language is not present in SHARE data. To have more control over the recoding process, we recoded job descriptions using CASCOT-NL in its manual mode with the assistance of a Dutch-native language team of researchers at SHARE partner CentERdata. As expected, disagreement rates with the topmost recommended code proposed by CASCOT were almost negligible for highly scored job descriptions. For instance, for the last job variable, only 10 job descriptions out of 968 to which CASCOT attributed a score higher than 80 were manually changed. Consequently, had we run CASCOT in semi-automatic mode setting a minimum score equal to 80 would have resulted in very similar codes.

Two main issues arise when comparing codes from SHARE and CASCOT-NL. The first one is the homogeneity of the classification structure. SHARE Netherlands coded job descriptions in *3-digit* ISCO-88 (Note that all other countries coded jobs in ISCO-88 at *4-digit* level, see above). CASCOT-NL codes, as described earlier, to ISCO-08 4-digit level. We then homogenised the two sets of codes as follows. First, we converted CASCOT-NL codes from ISCO-08 into ISCO-88 using official correspondence table (ILO, 2014). Unfortunately, there is no one-to-one correspondence between ISCO-08 and ISCO-88, i.e. multiple ISCO-88 codes are associated to the same 4-digit ISCO-08 code. In our data, this occurs for 220 individuals, i.e. 1/5 of the sample. In these cases, we associate multiple ISCO-88 codes to the same job description. Considering the issue of no one-to-one correspondence between different versions of ISCO, we state that a job description has a "different code" if the ISCO-88 code attributed by SHARE coders is not equal to *any* of the ISCO-88 codes resulting from the conversion into ISCO-88 of the CASCOT-NL output. Otherwise, we state that a job description has "same code". Second, we only consider 3-digits. To sum up, we compare codes from SHARE and CASCOT-NL in terms of 3-digit ISCO-88.

The second issue concerns coding vague and incomplete answers. As described earlier, SHARE coders and CASCOT-NL follow two different approaches for these types of job descriptions: whereas CASCOT-NL makes use of a separate category ("99.."), SHARE uses trailing zeros. As a result, vague and inadequate responses could not be compared, and are excluded from the statistical analysis. We also exclude those answers which were coded by CASCOT as "no conclusion".

Table 1 shows the sample size for our statistical analysis, i.e. 1,690 observations of which 1,083 concern last job and 607 current job. The higher frequency for last job in comparison with current job mostly reflects the distribution of respondents by work status in the first wave of SHARE.

Table 1: coding comparability in SHARE and CASCOT – Dutch data

	Last job		**Current job**	
	Freq.	*Percent*	*Freq.*	*Percent*
Comparable	1,083	62.1	607	60.82
Not comparable	661	37.9	391	39.18
Total	1,744	100	998	100

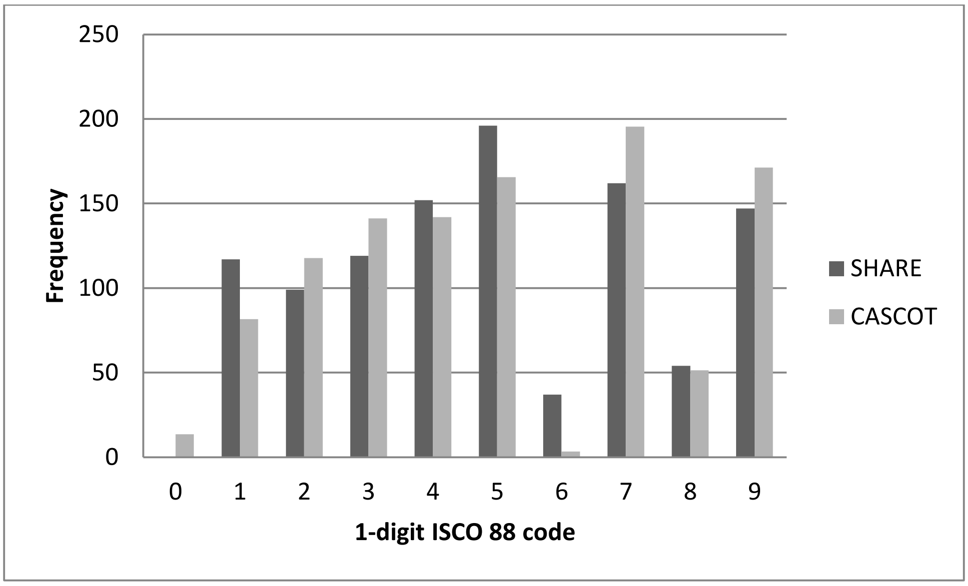
4. Results

4.1 Descriptive statistics

Figures 1a and 1b show the distribution of occupations by ISCO-88 major groups according to both SHARE and CASCOT-NL coding, for last and current job respectively. Given the fact that, due to the lack of one-to-one correspondence between ISCO-08 and ISCO-88, in our recoding exercise multiple codes are sometimes associated to the same individual, we use weights to construct these figures: In particular, when *n* codes are associated to the same individual, we attribute a weight equal to *1/n* to each of them.

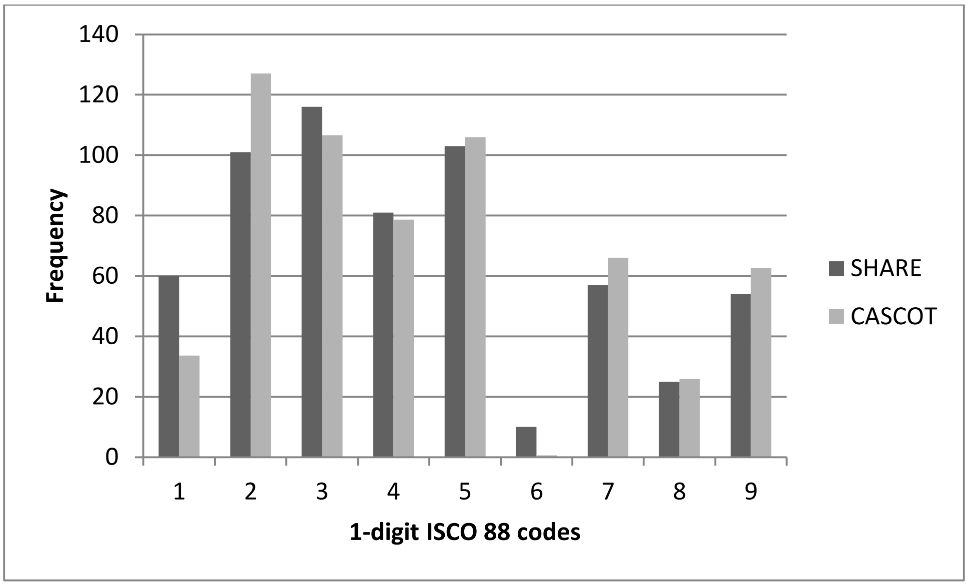
The figures highlight sizable differences between ISCO distributions of current and last job. The share of professionals and associate professionals (ISCO major groups 2 and 3) is much higher for current job than for last job, whereas the opposite occurs for lower-skilled occupations. This fact may reflect changes in the occupational structure over time, possibly due to technological change or international trade, as last job may easily refer to occupations started early in an individual's working career. There is in fact an extensive literature showing that technological progress and increased competition from low wage countries have changed labour demand in favour of more skilled occupations (e.g. Autor et al. 2003; Feenstra and Hanson 1996). In addition, these differences in the distribution of occupation can also be due to selective retirement: manual workers may retire earlier from the labor force than non-manual workers and therefore may be overrepresented in the last job variable; the contrary may occur for professionals, which may stay in the labor market even beyond the standard retirement age. The issue of selective retirement is non-negligible in countries favoring part-time work such as the Netherlands. Finally, note that the number of observations for each major group is limited; consequently, statistical analyses disaggregated by ISCO groups at 2/3-digits are not presented in this section.

Figure 1a: Distribution of occupation ISCO-88 major groups, CASCOT and SHARE coding – Last job



Legend: 0=Armed forces, 1=legislators, senior officials and manager, 2=professional, 3=technicians and associate professional, 4=clerks, 5=service workers and shop and market sale, 6=skilled agricultural and fishery workers, 7=craft and related trades workers, 8=plant and machine operators and assemblers, 9=elementary occupations

Figure 1b: Distribution of occupation ISCO-88 major groups, CASCOT and SHARE coding – Current job



Legend: 1=legislators, senior officials and manager, 2=professional, 3=technicians and associate professional, 4=clerks, 5=service workers and shop and market sale, 6=skilled agricultural and fishery workers, 7=craft and related trades workers, 8=plant and machine operators and assemblers, 9=elementary occupations

Tables 2a and 2b report frequency and percentage of same and different codes for last and current job respectively. The percentage of differently coded (which we call "disagreement rate" hereafter) appears high even when the comparison is made at 1-digit level (33.7 percent for last job and 40 percent for current job). As expected, such percentages rise with the number of digits the comparison is performed. Remarkably, the percentage of differently coded is sensibly higher for current job than for last job: e.g. at 3-digit level 60 percent of texts for current job are differently coded, cf. with 49 percent for last job. A possible explanation of this last finding is related to sample composition: we have seen that the ISCO-88 major group distribution for current and last job are sensibly different (Figure 1), and some ISCO groups may be more subject to coding errors than others (see Table 3). It has to be pointed out that previous exercises (Ellison, 2014) found qualitatively similar findings, namely when asked through open-ended questions mother's and father's jobs are typically better coded than individuals' own jobs. The intuition behind these results is that individuals tend to give too many details about their current job, because they think that their job is complex and do not provide easy descriptions, whereas this occurs to a lesser extent for parents' and last job.

Table 2a – same and different code: Last job

	1-digit		2-digit		3-digit	
ISCO-88 Code:	*Freq.*	*Percent*	*Freq.*	*Percent*	*Freq.*	*Percent*
same	718	66.3	639	59.0	548	50.6
different	365	33.7	444	41.0	535	49.4
Total	1,083	100	1,083	100	1,083	100

Table 2b – same and different code: Current job

	1-digit		2-digit		3-digit	
ISCO-88 Code:	*Freq.*	*Percent*	*Freq.*	*Percent*	*Freq.*	*Percent*
same	364	60.0	299	49.3	242	39.9
different	243	40.0	308	50.7	365	60.1
Total	607	100	607	100	607	100

Table 3 reports disagreement rates by ISCO-88 major groups, for both current and last job. There exists a wide heterogeneity in the disagreement rate across groups, with groups 1 ("legislators, senior officials and manager") and 3 ("technicians and associate professional") being those with the highest values. The percentage of differently coded is also high for the current job variable in group 6 ("skilled agricultural and fishery workers"). Agricultural workers are known to be difficult to code and some occupations in this category have been subjected to changes in classification from ISCO-88 to ISCO-08. The high disagreement rate for this category may be due to the fact that the ISCO-88 Unit groups 1221, "Production and operations department managers in agriculture forestry and fishing" and 1311, "General managers in agriculture forestry and fishing" have been removed from Major Group 1 in the ISCO 08-classification. The occupations included within this category have been moved to Sub-Major Group 61 and have been merged with the relevant supervisory groups (UN, 2007). Therefore, "General managers in agriculture hunting, forestry and fishing" are classified as ISCO-88 unit group 1311, and should not be included within group 6.

Table 3 – disagreement rate by ISCO major groups: last and current job (%)

	Last job			Current job		
	Disagreement rate (%)			Disagreement rate (%)		
ISCO 1-digit as coded in SHARE	**3-digit**	**2-digit**	**1-digit**	**3-digit**	**2-digit**	**1-digit**
Legislators, Senior Officials And Manager	82	65	59	80	53	47
Professional	44	37	34	38	31	28
Technicians And Associate Professional	64	52	50	70	59	53
Clerks	52	33	31	48	36	32
Service Workers And Shop And Market Sale	40	39	31	38	36	26
Skilled Agricultural And Fishery Workers	24	24	22	80	70	60
Craft And Related Trades Workers	30	20	09	61	35	16
Plant And Machine Operators And Assemblers	44	39	28	32	24	16
Elementary Occupations	39	25	17	72	56	31

In addition to disagreement rates, in the following we attempt to quantify the degree of disagreement between the two sets of codes. To do this, we need to assume that the order of ISCO-88 major groups, from "1" to "9" (while Armed forces are not part of this ordering), is meaningful. To be clearer, a job description *x* is considered to be more differently coded than a job description *y* if the former is e.g. coded as "1" in SHARE and as "9" in CASCOT, while the latter is e.g. coded as "1" in SHARE and as "2" in CASCOT. Considering the issue of no one-to-one correspondence between different versions of ISCO (see above), we use weights when constructing bivariate distributions in Table 4a and 4b (e.g. if we obtain 3 possible ISCO-88 codes for a given job description, we attribute a weight equal to 1/3 to each of them). We first perform the Wilcoxon signed-rank test for paired data (Wilcoxon 1945). The null hypothesis that SHARE and CASCOT-NL coding distributions are the same is rejected at 0.5% confidence level for last job and at 4.3% level for current job.

The bivariate distributions – SHARE vs CASCOT-NL ISCO-88 major groups – are presented for last job in Table 4a and for current job in Table 4b. The percentages reported in these tables sum up to 100 percent horizontally, i.e. with respect to SHARE coding. For instance, 41.5 percent of job descriptions coded as "1" ("legislators, senior officials and manager") by SHARE coders have also been coded as "1" by CASCOT-NL, while the same software has coded about 13 percent of them as "2" ("professionals"). Despite the low frequency of observations, which may limit the statistical validity of some of these figures, the off-main diagonal cells of these matrixes probably highlight some common coding problems. One of them is the remarkable percentage of 55.6 percent (Table 4a, 1st column, 6th row) coded in group 1 by CASCOT and in group 6 by SHARE, which likely reflects the difficulty in coding "General managers in agriculture, hunting, forestry and fishing" (CASCOT performs better than SHARE in this case if this is true). This result should be taken with caution considering the very low number of observations in our sample for this group (N=10 for current job and N=37 for last job). However, what is reassuring is that most of the coding disagreement occurs within similar groups of occupations (1 to 3, 4 to 7, and 8 to 9), which means that if occupations are used to construct social class indices (see for example Harrison, 2010), the classification errors should not be too pronounced.

Table 4a – Bivariate distributions - SHARE vs CASCOT-NL ISCO-88 major groups - Last job (%)

Cascot Share ↓	1	2	3	4	5	6	7	8	9	Total
1	41.6	13.3	22.7	7.0	3.5	0.0	9.8	0.7	1.4	100
2	1.2	63.1	27.0	3.5	3.5	0.0	0.6	1.2	0.0	100
3	7.0	29.6	44.5	5.1	7.2	0.5	3.1	0.0	3.1	100
4	1.1	4.2	18.8	69.7	1.4	0.0	1.4	0.0	3.5	100
5	11.3	1.7	5.0	0.6	70.5	0.0	3.3	3.3	4.4	100
6	55.6	0.0	0.0	0.0	0.0	8.9	4.4	0.0	31.1	100
7	0.0	0.0	1.3	0.0	0.7	0.0	90.3	3.9	3.9	100
8	0.0	2.1	4.9	4.2	0.0	0.0	16.9	63.4	8.5	100
9	0.4	1.4	1.4	6.9	2.4	0.4	4.2	1.4	81.7	100
Total	7.6	11.0	13.2	13.3	15.5	0.3	18.3	4.8	16.0	100

Legend: 1=legislators, senior officials and manager , 2=professional, 3=technicians and associate professional, 4=clerks, 5=service workers and shop and market sale, 6=skilled agricultural and fishery workers, 7=craft and related trades workers, 8=plant and machine operators and assemblers, 9=elementary occupations

Table 4b – Bivariate distributions - SHARE vs CASCOT-NL ISCO-88 major groups – Current job (%)

Cascot → Share ↓	1	2	3	4	5	6	7	8	9	Total
1	43.5	20.9	13.0	4.9	2.5	0.0	12.3	0.5	2.5	100
2	1.0	69.9	18.3	3.6	6.0	0.0	1.2	0.0	0.0	100
3	2.4	38.8	40.3	3.9	8.7	0.0	1.0	0.0	4.9	100
4	0.0	2.5	25.6	68.1	3.8	0.0	0.0	0.0	0.0	100
5	6.6	1.1	8.8	0.0	76.8	0.0	1.1	1.1	4.4	100
6	29.3	0.0	4.9	0.0	0.0	7.3	0.0	0.0	58.5	100
7	0.0	0.0	2.8	0.0	1.9	0.0	85.1	4.7	5.6	100
8	0.0	0.0	0.0	0.0	0.0	0.0	14.9	76.6	8.5	100
9	1.0	3.8	2.9	11.4	8.6	0.0	1.9	1.9	68.6	100
Total	6.1	17.0	22.2	12.0	15.8	2.1	9.4	5.4	9.9	100

Legend: 1=legislators, senior officials and manager, 2=professional, 3=technicians and associate professional, 4=clerks, 5=service workers and shop and market sale, 6=skilled agricultural and fishery workers, 7=craft and related trades workers, 8=plant and machine operators and assemblers, 9=elementary occupations

The ILO maps ISCO major groups into skill levels (Elias 1997; ILO 2012) which can be then mapped to ISCED-97 levels of education (see Table A1 in the Appendix). Tables 5a and 5b present the bivariate distributions – SHARE vs CASCOT-NL skill levels groups - for respectively last and current job. The tables confirm that most of the coding disagreement occurs within similar groups of occupations. When grouping occupations according to their skill level, we note that the percentages of occupations that are coded in the same skill group is reasonably high. Looking at last job, 82% of occupations coded in skill group 1 in SHARE are coded in the same group in CASCOT as well. The percentages of correct coding are around 80% for skill group 2, 57% for skill group 3 and 63% for skill group 4. As seen before, these percentages are lower when considering current job.

Table 5a – bivariate distributions - SHARE vs CASCOT-NL skill levels - Last job (%)

Cascot → Share ↓	1	2	3	4	Total
1	81.66	15.22	1.73	1.38	100.00
2	5.46	79.11	13.61	1.82	100.00
3	2.36	18.02	56.88	22.74	100.00
4	0.00	8.76	28.19	63.05	100.00
Total	16.01	52.15	20.83	11.01	100.00

Table 5b – bivariate distributions - SHARE vs CASCOT-NL skill levels - Current job (%)

Cascot → Share ↓	1	2	3	4	Total
1	68.57	23.81	3.81	3.81	100.00
2	5.12	78.65	15.05	1.18	100.00
3	4.18	15.45	46.61	33.76	100.00
4	0.00	10.85	19.25	69.90	100.00
Total	10.32	45.65	23.10	20.93	100.00

In the remaining part of the article, we investigate which individual characteristics are more likely associated to different coding. We perform both univariate and multivariate analyses. We show tables reporting univariate statistics in the Appendix. In particular, Table A2 shows the disagreement rate by education, Table A3 by gender and Table A4a and A4b by industry for last and current job respectively. The figures clearly show that the rates of coding disagreement differ substantially across education and gender, with higher rates for more educated individuals (only for last job) and for males. No clear patterns emerge from the tables on disagreement rates by industry, probably because of the very low number of observation in some groups. In the next subsection, we investigate this result in more details by performing a multivariate analysis.

4.2 Multivariate analysis

What individual characteristics are associated to the probability of having provided an answer to the SHARE question "what is your [main/last] job called? Please give the exact name or title" which has been differently coded in SHARE and CASCOT-NL? Among these characteristics, we specifically explore the role of education and gender, but we also shed some light on the importance of two basic job-related characteristics (industry and ISCO group) on the probability of coding disagreement.

We estimate a set of linear probability models (LPM) for coding disagreement. A LPM is a multiple linear regression model with a binary dependent variable (Wooldridge 2010). The dependent variable of these models allows for the possibility of multiple correspondences in the ISCO-08 to ISCO-88 conversion tables. In other words, in our models the dependent variable is a dummy variable equal to 1 if the ISCO-88 code provided by SHARE is not equal to any of the ISCO-88 codes resulting from the conversion into ISCO-88 of the ISCO-08 CASCOT code; otherwise, the dependent variable is equal to 0. We consider three types of the dependent variable, depending on the number of digits at which we compare SHARE and CASCOT codes, namely a dummy for being differently coded at 1-digit, at 2-digits, or at 3-digits.

The set of LPM we estimate differ in terms of the dependent variable as explained above, and in terms of the set of explanatory variables. We estimate separate models for current and for last job. By pooling together these two variables, we would have considerably increased the number of observations and perhaps improved the precision of our estimates. Nevertheless, the descriptive findings outlined earlier suggest that coding disagreement for current and last job follows different patterns; our econometric results (see later) clearly confirm that pooling current and last job together – assuming that explanatory variables have same effect on the probability of different coding for current and last job - would have led to mis-specification.

Table 5a reports LPM estimates for the probability of the last job to be differently coded at 3-digit level. We present four specifications in this table. Specification (1) includes dummy variables for gender and educational attainment (four aggregated ISCED-97 groups) as explanatory variables. Our results indicate that females show a 20 percent lower probability to be differently coded when compared to males. Remarkably, we also found that there is a strong positive gradient between education and coding disagreement: relative to individuals with no or primary education, those with a lower-secondary degree (ISCED 2) have a 10 percent higher probability of different coding; this percentage raises to about 17 percent for individuals with an upper and post-secondary degree (ICED 3-4), up to 28 percent for those holding a tertiary education degree (ISCED 5-6).

These results are particularly interesting, as they suggest that the probability of being miscoded is not random, but is more pronounced for certain groups. In particular, it seems that more educated individuals and males are more likely to be coded differently when using alternative coding systems. This may be due the fact that males and more educated people are sorted in particular occupations that are intrinsically more difficult to be classified. In fact, for example, more educated individuals and males are likely to work in high skilled occupations - as shown for the mean level of education and the percentage of females for each 1-digit group ISCO-88 in Table A5 in the appendix - where the coding disagreement is higher according to the results shown in Table 3. An alternative explanation could be that education and gender affect somehow the way people are able to describe their jobs when asked in interviews.

Specification (2) adds two right-hand-side variables to the model. A dummy for being coded as "not elsewhere classified (NEC)" was constructed by looking at the ISCO-88 4-digit codes, as coded by CASCOT software. This dummy is equal to 1 if the ISCO-88 fourth digit is equal to 9, which, according to ILO's guidance, refers to occupational categories that are not classified to other specific categories within the classification. This variable includes ISCO categories, which usually contain many types of clerical jobs. We thus expect NEC jobs to be more likely differently coded. More important, since these jobs are typically performed by females, including this variable is expected to affect the estimate for the gender variable. Another dummy was constructed for the self-employed. Being self-employed is also correlated with gender. As expected, the variable "not elsewhere classified" is positive and significant at 10 percent level; however, the coefficient for females is not affected by controlling for this confounding factor. The self-employed variable turned out to be not significant.

In specification (3) we additionally control for industry by including in the model a set of 31 industry dummy variables. Industry is classified using NACE Codes, Version 4 Rev. 1 1993 (see http://www.top500.de/nace4-e.htm for a description of NACE Version 4 Rev. 1 and MEA, 2013, pp. 32-33 for the shorter classification used in SHARE). They jointly affect the probability of different coding, as indicated by the result of the Wald test reported at the bottom of the table (p-value 0.02). Once controlling for industry, the positive gradient between coding disagreement and education attainment shown in the previous specifications becomes less clear: only the tertiary education dummy variable remains strongly significant. Moreover, the coefficient for female reduces in size (from -.20 to -.15).

Specification (4) builds upon specification (3) by adding to it a full set of ISCO 3-digit dummy variables (90 groups). This specification is very demanding in terms of data requirements, and we expect to have limited variability in gender and, especially, in education once we condition on being coded in a specific ISCO minor group. The most clear-cut effect of adding ISCO unit groups to the model is the dramatic increase in the model fit: the R^2 (see the ancillary statistics at the bottom of the table) in fact increases from about 12 percent (specification c) to about 44 percent (specification d). The p-value of the Wald test for no joint significance of the ISCO minor groups dummy variables is equal to 0. Controlling for ISCO minor groups determines a sizable reduction in the coefficient for female (from -.15 in specification 3 to -.1 in specification 4). Adding ISCO minor groups has an overall quite limited impact on the coefficients for education: the dummy variable for having attaining a Tertiary education degree (ISCED 5-6) is equal to .16 (cf. with .24 in specification 3) and remains highly significant. These last findings remain almost unchanged if we condition on either ISCO 2-digit or ISCO 1-digit groups instead of ISCO 3-digit groups.

Table 5a – LPM for the probability to be differently coded at 3-dgt level: estimation results, last job

VARIABLES	(1)	(2)	(3)	(4)
Female	-0.205***	-0.207***	-0.152***	-0.101**
	(0.030)	(0.030)	(0.038)	(0.040)
Lower-secondary education (ISCED 2)	0.100***	0.098**	0.060	-0.009
	(0.038)	(0.038)	(0.043)	(0.038)
Upper and post-secondary education (ISCED 3-4)	0.168***	0.168***	0.095*	0.031
	(0.045)	(0.045)	(0.050)	(0.047)
Tertiary education (ISCED 5-6)	0.280***	0.276***	0.236***	0.160***
	(0.052)	(0.052)	(0.060)	(0.060)
Not elsewhere classified		0.147*	0.048	-0.082
		(0.082)	(0.093)	(0.086)
Self-employed		-0.079	-0.052	-0.014
		(0.052)	(0.060)	(0.056)
Additional controls:				
Industry dummy (31 groups)	No	No	Yes	Yes
ISCO 3-digit dummy (90 groups)	No	No	No	Yes
Ancillary statistics:				
Wald test H0: no joint significance industry dummy variables (p-value)	-	-	0.0213	0.0203
Wald test H0: no joint significance ISCO 3-digit dummy variables (p-value)	-	-	-	0
Observations	1,066	1,066	933	933
R-squared	0.079	0.083	0.117	0.443

Note: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; Reference categories: male, no or primary education (ISCED 0-1), employee.

Table 5b reports LPM estimates for the probability of the current job to be differently coded at 3-digit level. To facilitate comparability, we report the same four specifications presented in Table 5a. Results for the current job are very different from those obtained for the last job: female is negatively associated to coding disagreement in specifications (1) to (3) while this coefficient loses its significance once controls for ISCO minor groups are added to the model (specification 4). There is no education coding disagreement gradient for the current job variable. Industry and ISCO minor groups maintain their strong explanatory power (see results of corresponding Wald tests at the bottom of the table).

Finally, we point out that results for both last and current job variable remain almost unchanged if we change the dependent variable from coding disagreement at 3-digit level to disagreement at 1-

or 2-digit levels; these results are available from the authors upon request. They are also unaffected if we run CASCOT in semi-automatic mode instead of in its one-by-one mode.

Table 5b – LPM for the probability to be differently coded at 3-dgt level: estimation results, current job

VARIABLES	(1)	(2)	(3)	(4)
Female	-0.140***	-0.140***	-0.083*	-0.020
	(0.041)	(0.041)	(0.048)	(0.050)
Lower-secondary education (ISCED 2)	-0.035	-0.031	-0.017	-0.046
	(0.084)	(0.084)	(0.088)	(0.085)
Upper and post-secondary education (ISCED 3-4)	-0.055	-0.056	-0.023	-0.132
	(0.086)	(0.086)	(0.092)	(0.091)
Tertiary education (ISCED 5-6)	-0.035	-0.031	0.027	-0.154
	(0.084)	(0.084)	(0.094)	(0.097)
Not elsewhere classified		0.058	0.038	-0.057
		(0.102)	(0.103)	(0.106)
Self-employed		-0.074	-0.004	-0.035
		(0.058)	(0.065)	(0.068)
Additional controls:				
Industry dummy (31 groups)	No	No	Yes	Yes
ISCO 3-digit dummy (90 groups)	No	No	No	Yes
Ancillary statistics:				
Wald test H0: no joint significance industry dummy variables (p-value)	-	-	0.0089	0.0065
Wald test H0: no joint significance ISCO 3-digit dummy variables (p-value)	-	-	-	0
Observations	602	602	531	531
R-squared	0.020	0.024	0.113	0.439

Note: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; Reference categories: male, no or primary education (ISCED 0-1), employee.

5. Conclusions

This article studied the potential measurement errors occurring when coding occupational data. Given the growing use of information on occupation in labour economics research, the quality of occupational data is of key importance and is often neglected by the economic literature.

In this analysis, we have recoded open-ended questions on occupation for the Dutch sample of SHARE data using CASCOT, a well-known software for automatic ex-post coding. Our results show that the disagreement rate, defined as the percentage of observations coded differently in SHARE and CASCOT, is high even when the comparison is made at 1-digit level (33.7 percent for last job and 40 percent for current job). This finding is particularly striking, considering that our approach has been conservative, in the sense that we only compare the “easiest” answers, because vague and incomplete answers are left out form the analysis. The level of miscoding we find should thus be considered as a lower bound of the “true” miscoding.

In our view our results highlight the complexity of occupational coding and suggest that the potential measurement error due to miscoding should be taken into account when making statistical analysis or writing econometric models.

We have also tested whether such a measurement error in occupation is random or is instead correlated to some specific individual or job-related characteristics. We found that the measurement error is indeed more evident in certain ISCO-88 groups (ISCO-88 groups 1 and 3) and is more pronounced for more educated individuals and males. This may be due to the fact that males and more educated people are sorted in particular occupations that are intrinsically more difficult to be classified. Alternatively, it could be that education and gender affect somehow the way people are able to describe their jobs when asked in interviews. Understanding the reasons behind these results may constitute an interesting direction for further investigation.

References

Autor, D. 2013. "The "task approach" to labour markets: an overview." *Journal of Labour Market Research* 46: 185-199.

Autor D, L F Katz, and M S Kearney. 2006. "The Polarization of the US Labor Market." *American Economic Review* 96: 189-194.

Autor, D., F. Levy and R.J. Murnane. 2003. "The Skill Content Of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118: 1279-1333.

Bethmann A., M. Schierholz, K. Wenzig, and M. Zielonka. 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.

CBS (Statistics Netherlands), 2012. *Coding tool implemented in 2012 for coding occupations in social surveys, internal document*. The Hague: Statistics Netherlands.

Cheeseman Day J. 2014. *Using an Autocoder to Code Industry and Occupation in the American Community Survey*. Presentation for the Federal Economic Statistics Advisory Committee Meeting June 13, 2014.

Commission of the European Communities. 2009. "Commission Regulation (EC) No 1022/2009 of 29 October 2009 amending Regulations (EC) No 1738/2005, (EC) No 698/2006 and (EC) No 377/2008 as regards the International Standard Classification of Occupations (ISCO)". *Official Journal of the European Union*, L 283/3, 30.10.2009.

De Vries J. and H.B.G. Ganzeboom. 2008. "Hoe meet ik beroep? Open en gesloten vragen naar beroep toegepast in een statusverwervingsmodel." *Mens & Maatschappij* 83: 71-96 plus "Rectificatie." *Mens & Maatschappij* 83: 190-191.

Elias, P., K. Halstead and K. Prandy 1992. *CASCOC: Computer-assisted Standard Occupational Coding*. London: HMSO.

Elias, P. 1997. "Occupational Classification (ISCO-88. Concepts, Methods, Reliability, Validity and Cross-National Comparability". *OECD Labour Market and Social Policy Occasional Papers*, No. 20, OECD Publishing. http://dx.doi.org/10.1787/304441717388.

Ellison, R. 2014. *Demonstration of performance of CASCOT 5.0*. Presentation at the *CASCOT:* Occupational Coding in Multi-national Surveys Workshop in Venice (10-11 April 2014), http://dasish.eu/dasishevents/cascotworkshop/programmepres/.

Feenstra, R. C. and G. H. Hanson. 1996. "Globalization, Outsourcing, and Wage Inequality." *American Economic Review* 86: 240-245.

Fletcher, J. M., J. L. Sindelar and S.Yamaguchi. 2011. "Cumulative effects of job characteristics on health." *Health Economics* 20: 553-570.

Ganzeboom, H. 2008. Occupation Coding: Do's And Dont's. Version 2, August 2 2008. Available at www.gesis.org.

Goos, M. and A. Manning. 2007. “Lousy and lovely jobs: The rising polarization of work in Britain.” *Review of Economics and Statistics* 89: 118-133.

Harrison, E. 2010. *Social Class in Europe: An Introduction to the European Socio-economic Classification*. London: Routledge.

Hartog, J. 2000. “Over-education and earnings: where are we, where should we go?” *Economics of education review* 19: 131–147.

Hoffmann, E., P. Elias, B. Embury and R Thomas. 1995. *What Kind Of Work Do You Do? Data collection and processing strategies when measuring "occupation" for statistical surveys and administrative records*. Geneva: ILO working paper, N.95-1.

Jackle, A. 2008. “Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection.” *Journal of Official Statistics* 24: 411-430.

Jones, R. and P. Elias 2004. 'CASCOT: Computer-assisted Structured Coding Tool', Coventry: Warwick Institute for Employment Research, University of Warwick.

ILO, 2014. ISCO: International Standard Classification of Occupations. Available at: http://www.ilo.org/public/english/bureau/stat/isco/

ILO, 2012. *International Standard Classification of Occupations: Structure, group definitions and correspondence tables*. Geneva: Volume 1.

ILO, 2010. *Measuring the Economically Active in Population Censuses: A Handbook*. New York: ILO and UN, Studies in Methods Series F, No. 102.

Leuven, E. and H. Oosterbeek. 2011. Overeducation and mismatch in the labor market, in E. Hanushek, S. Machin & L.Woessmann, eds, ‘*Handbook of the Economics of Education*’, Vol. 4, Elsevier.

MEA 2013. SHARE Release Guide 2.6.0. Waves 1 & 2. Munich.

Ravesteijn, B., H. van Kippersluis and E. van Doorslaer. 2013. *The Wear and Tear on Health: What is the Role of Occupation?*. Amsterdam: Tinbergen Institute Discussion Paper 13-143.

Tijdens, K.G. 2014a. “Drop-out rates during completion of an occupation search tree in web-surveys”. *Journal of Official Statistics* 30: 23–43.

Tijdens, K.G. 2014b. *Reviewing the measurement and comparison of occupations across Europe*. Amsterdam: University of Amsterdam, AIAS Working Paper 149.

United Nations, 2007. Updating the International Standard Classification of Occupations (ISCO): Summary of major changes between ISCO-88 and ISCO-08 (Feb 2007 draft), Meeting of the Expert Group on International Economic and Social Classifications, New York, 16-18 April 2007.

United Nations, 2014. National Classifications. Available at: http://unstats.un.org/unsd/cr/ctryreg/ctrylist2.asp

Westerman, S. 2014. *CBS and CASCOT: tuning CASCOT for improved performance*. Presentation at the *CASCOT:* Occupational Coding in Multi-national Surveys Workshop in Venice (10-11 April 2014), http://dasish.eu/dasishevents/cascotworkshop/programmepres/.

Wilcoxon, F. 1945. "Individual comparisons by ranking methods." *Biometrics* 1: 80–83.

Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge (MA): MIT press.

Appendix

Table A1: Mapping of ISCO-08 major groups to skill levels (col. 1 and 2) and mapping of the four ISCO-08 skill levels to ISCED-97 levels of education (col. 2 and 3)

ISCO-08 major groups	Skill level	ISCED-97 level
1. Managers	3 + 4	5b + 6, 5a
2. Professionals	4	6, 5a
3. Technicians and associate professionals	3	5b
4. Clerical support workers	2	4, 3, 2
5. Services and sales workers	2	4, 3, 2
6. Skilled agricultural, forestry and fishery workers	2	4, 3, 2
7. Craft and related trades workers	2	4, 3, 2
8. Plants and machinery operators, and assemblers	2	4, 3, 2
9. Elementary occupations	1	1

Note: ISCED-97 levels of education: Level 1=Primary education or first stage of basic education; Level 2 = Lower secondary or second stage of basic education; Level 3 = (Upper) secondary education; Level 4 = Post-secondary non-tertiary education; Level 5a = First stage of tertiary education, 1st degree, medium duration; Level 5b= First stage of tertiary education , short or medium duration, practical orientation); Level 6 = Second stage of tertiary education.

Source: ILO (2012), p. 14

Table A2: Disagreement rate by education levels

	Last job				Current job			
		Disagreement rate (%)				Disagreement rate (%)		
	N	3-digit	2-digit	1-digit	*N*	3-digit	2-digit	1-digit
ISCED 0-1	237	35	27	20	42	60	52	38
ISCED 2	465	44	34	27	208	55	42	29
ISCED 3-4	227	53	42	37	155	54	43	34
ISCED 5-6	137	67	54	49	197	55	41	37
Total	1066	47	37	31	602	55	43	34

Table A3: Disagreement rate by gender

	Last job				Current job			
		Disagreement rate (%)				Disagreement rate (%)		
	N	3-digit	2-digit	1-digit	*N*	3-digit	2-digit	1-digit
Male	536	59	46	38	332	61	45	35
Females	547	36	28	24	275	48	40	32
Total	1083	47	37	31	607	55	43	34

Table A4a: Disagreement rate (%) by industry (NACE codes) – last job (sorted by disagreement rate at 3-digit)

Industry	N	Disagreement rate (%) 3-digit	2-digit	1-digit
Computer and related activities	1	100	100	100
Recycling	1	100	100	100
Real estate activities, Renting of machinery and equipment without operator and of personal and household goods	5	80	40	40
Manufacture of coke, refined petroleum products and nuclear fuel	9	78	78	78
Electricity, gas, steam and hot water supply	16	75	56	38
Research and development	4	75	75	50
Publishing, printing and reproduction of recorded media	23	74	70	70
Education	50	72	54	42
Wholesale trade and commission trade, except of motor vehicles and motorcycles	26	69	62	54
Manufacture of basic metals, metal products except machinery & equipment	19	63	63	53
Financial services and Insurance	21	62	19	19
Manufacture of other non-metallic mineral products	5	60	60	40
Other business activities	47	60	47	38
Transport, Post, Telecommunications	53	58	51	34
Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	7	57	29	29
Hotels and restaurants	18	56	56	17
Manufacture of motor vehicles, trailers and semi-trailers	9	56	44	44
Public administration and defence; compulsory social security	92	53	45	41
Recreational, cultural and sporting activities	23	52	43	39
Mining	53	51	47	40
Manufacture of food, tobacco, textiles, clothes, bags, leather goods	64	50	47	38
Sewage and refuse disposal, sanitation and similar activities	2	50	0	0
Construction	95	47	37	26
Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel	12	42	42	25
Manufacture of electronic or electric machinery and devices	5	40	20	20
Health and social work	126	39	31	28
Activities of membership organization n.e.c.	15	33	20	20
Other service activities	34	32	32	29
Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods	105	31	30	25
Manufacture of furniture; manufacturing n.e.c.	4	25	25	25
Manufacture of machinery and equipment n.e.c.	4	0	0	0
Total	948	50	42	34

Note: Industry is classified using NACE Codes, Version 4 Rev. 1 1993 (see http://www.top500.de/nace4-e.htm for a description of NACE Version 4 Rev. 1 and MEA, 2013, pp. 32-33 for the shorter classification used in SHARE).

Table A4b: Disagreement rate (%) by industry (NACE codes) – current job (sorted by disagreement rate at 3-digit)

		Disagreement rate (%)		
Industry	**N**	**3-digit**	**2-digit**	**1-digit**
Manufacture of motor vehicles, trailers and semi-trailers	3	100	67	33
Research and development	1	100	100	100
Mining	23	87	74	57
Other business activities	39	85	74	67
Education	59	81	66	54
Real estate activities, Renting of machinery and equipment without operator and of personal and household goods	5	80	40	40
Electricity, gas, steam and hot water supply	4	75	50	25
Hotels and restaurants	8	75	75	25
Construction	43	72	49	42
Manufacture of food, tobacco, textiles, clothes, bags, leather goods	13	69	62	38
Computer and related activities	9	67	67	33
Manufacture of basic metals, metal products except machinery & equipment	3	67	67	33
Recreational, cultural and sporting activities	20	65	60	30
Transport, Post, Telecommunications	26	62	50	42
Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	5	60	60	60
Financial services and Insurance	12	58	50	50
Public administration and defence; compulsory social security	52	58	54	42
Wholesale trade and commission trade, except of motor vehicles and motorcycles	2	50	50	50
Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods	32	47	44	28
Health and social work	133	45	41	38
Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel	7	43	43	14
Manufacture of coke, refined petroleum products and nuclear fuel	3	33	0	0
Manufacture of electronic or electric machinery and devices	3	33	33	33
Publishing, printing and reproduction of recorded media	6	33	33	33
Manufacture of machinery and equipment n.e.c.	4	25	25	0
Other service activities	14	21	14	14
Activities of membership organization n.e.c.	5	20	20	20
Manufacture of furniture; manufacturing n.e.c.	2	0	0	0
Total	536	60	52	41

Note: Industry is classified using NACE Codes, Version 4 Rev. 1 1993 (see http://www.top500.de/nace4-e.htm for a description of NACE Version 4 Rev. 1 and MEA, 2013, pp. 32-33 for the shorter classification used in SHARE).

Table A5: Educational attainment and gender composition across ISCO-88 1 digit groups

ISCO 1-dgt	% primary	% lower secondary	% upper secondary	% tertiary	Mean years of education	% of female
1	5.6	30.4	29.9	34.1	14.0	20.3
2	0.8	14.2	21.2	63.7	16.1	54.6
3	3.2	22.8	35.1	38.9	14.0	41.5
4	7.8	50.4	32.6	9.2	12.6	72.4
5	18.9	54.7	21.6	4.8	11.6	81.9
6	20.0	61.4	12.9	5.7	11.2	42.3
7	31.5	48.2	17.5	2.8	9.8	20.6
8	29.8	49.7	17.1	3.3	10.9	20.0
9	35.3	50.5	10.7	3.6	9.9	70.6
Total	15.1	40.2	23.7	21.0	12.5	51.2

Note: The table is computed pooling current and last job and using SHARE coding