Automatic coding system of occupation of Statistics Netherlands

Sue Westerman, 28-06-2018

Contents

1. Introduction

Statistics Netherlands has the possibility to code responses to questions on occupation and the main tasks in social surveys according to the international classification of occupation ISCO 2008 (International Standard Classification of Occupations 2008). During the coding process the enormous variety of possible descriptions of occupation is structured and provided with codes of standard classifications of occupations. The information on occupation collected via surveys is then suitable for analysis.

This memo illustrates how the process of coding occupation is structured[1]. Occupation is coded by means of two distinct processes. The difference lies in the required level of detail and quality and the related proportion that is automatically or manually coded. One process variant focuses on delivering the 4-digit ISCO 2008 of sufficient quality in an optimal balance between automatic and manual coding, in the other process variant the proportion of manual coding is determined by the desired quality and detail. In the first chapter these two process variants are explained in further detail and differences in the starting points of the two processes are discussed. In the two following chapters, the process steps per variant, the input used, the output that is delivered and the quality thereof are discussed in detail for each of the process variants. In chapter 5 some insight is given into the way the coding tool was implemented in the first phase of development of the coding processes. Appendix 1 deals with the measurement of quality when the occupation is fully automatically coded, appendix 2 and 3 show the quality and distribution after fully automatic coding across the categories of two classifications of occupation, the International Standard Classification of Occupations 2008 (ISCO 2008)[2] and ROA-CBS 2014 (BRC 2014)[3] which is a national classification derived from the ISCO 2008 unit groups developed by Statistics Netherlands in close cooperation with Maastricht university. A crosstab in appendix 4 provides insight into the distribution of the coding on the highest aggregation level of the two classifications for both coding processes. Appendix 5 illustrates the progress of the two process variants by means of a schematic representation and a number of practical examples.

[1] With thanks to Elena Grigorieva, Birgit van Gils, Martijn Souren, Hendrika Lautenbach who contributed to the realization of the documentation by giving comments and asking questions. Thanks to Roel Schaart for translating the Dutch version of the documentation into English.

[2] http://www.ilo.org/public/english/bureau/stat/isco/isco08/

[3] https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs-en-beroepen/beroepenclassificatie--isco-en-sbc--

2. Coding process, two variants

For the purpose of coding information collected in social surveys, two process variants each with its own program sequence were designed.

In the first two steps of both processes coding takes place automatically on the basis of occupation and the combination of occupation and tasks. The relevant distinction between the two processes is the extent of additional help information on occupation that is used during automatic coding and the way manual coding is applied.

The two variants were not developed simultaneously and they meet specific needs for measuring the variable occupation. The coding process first developed aims to deliver adequate coding on the basis of the International Standard Classification of Occupations 2008 (ISCO 2008) at the highest level of detail with sufficient quality. This process is partly automatic and partly manual. The second coding process is more flexible. This variant allows for fully automatic coding and the use of manual coding is optional. This process aims to provide just sufficient quality at the requested level of aggregation for the purpose of publication and analysis.

The first variant of the coding process was developed during the redesign of the social surveys that took place in 2011 and 2012. The general objective of the redesign was to organize social surveys more cost-effectively. Accordingly, the coding process of occupation was expected to comply with the following requirements:

- The coding process had to be appropriate for the observation of data on occupation collected through three different interview modes: CAPI, CATI and CAWI, in the same manner in the 3 different modes and in different studies (LFS, NEA, ZEA, AKO etc.). During the redesign CAWI was for the first time introduced in the observation of occupation via social surveys.
- The coding process had to merely use generic software tools so as to save on development costs, costs of maintenance, management of customized tools and should preferably be suitable for both manual and automatic coding of occupation. This resulted in the choice of the Cascot tool developed by the University of Warwick[4].
- The national output requirement was to provide the ISCO at 4 digits.
- The coding efficiency of 4-digit automatic coding with sufficient quality should be at least equivalent but preferably higher to what the previous coding process[5] generated during automatic coding of occupations, which was at least 60%.

The second variant of the coding process was developed in 2016 for a customized assignment[6] for a research agency. It was arranged in such a way that the process would be more widely applicable and

[4] https://warwick.ac.uk/fac/soc/ier/software/cascot/

[5] See Computer assisted coding by interviewers, Wim Hacking, John Michiels, Saskia Janssen-Jansen, CBS 2008

[6] In 2015 this research agency asked Statistics Netherlands to develop a process with which the collected information on occupation (and tasks) of an survey among 160 thousand respondents can be fully automatically coded according to the International Standard Classification of Occupations 2008 (ISCO 2008). In 2016 this coding process was completed and data coded through this process were delivered in accordance with the custom request.

was put into use by Statistics Netherlands in 2018. This process had to comply with other requirements:

- It should be possible to provide answers to questions on occupation and/or tasks with valid ISCO codes fully automatically, even if not observed through the standard research question on occupation of Statistics Netherlands.
- For each category supplied data should be provided with an indication of the quality from the 1st to 3rd aggregation level of ISCO 2008 and BRC 2014.
- A flexible deployment of manual coding should be optional, namely only in case there is a need to improve the quality of the 1st, 2nd or 3rd aggregation level.

The choice for one or other variant depends on the client's output requirements. If a client wishes to use coded data of sufficient quality on occupation at the most detailed 4th aggregation level of ISCO 2008 or if occupation is a target variable, the most advanced variant of the coding process described (variant 1) is recommended. For this purpose it is necessary to use the Statistics Netherlands' standard question on occupation, see section 3.2. Then optimum use can be made of additional variables and manual coding capacity can be deployed for an acceptable quality level of 4-digit coding. If less detail is required for analysis purposes or if occupation is not a target variable but a background variable, the second process variant will suffice and saves costs. In theory, it is then possible to use questions on occupation different from the standard questions on occupation of Statistics Netherlands, see section 4.2.

Table 1 shows the aggregation levels, the number of digits of the coding per aggregation level and the number of corresponding categories of ISCO 2008 and BRC 2014 presented side by side. Both classifications have a hierarchical structure. In ISCO 2008 the highest aggregation level (major groups) consists of a 1 digit code, each lower aggregation level always has 1 additional digit. BRC 2014 consists of 2 digits at the highest level (occupational classes), 3 digits for the 2nd aggregation level and at the 3rd aggregation level 4 digits. The BRC 2014 occupations are identical to the unit groups of ISCO 2008 and have 4 digits. Sections 3.3 and 4.3 clarify for both process variants how the output is made available.

Table 1 The aggregation levels of ISCO 2008 and BRC 2014, number of groups, number of digits and the recommended process variant to be used per aggregation level.

Aggregation level	ISCO 2008	Number of groups	Number of digits	BRC 2014	Number of groups	Number of digits	Recommended process variant
1	Major groups	10	1 dig	Occupational classes	13	2 dig	Variant 2
2	Sub-major groups	43	2 dig	Occupational segments	41	3 dig	Variant 2
3	Minor groups	130	3 dig	Occupational groups	114	4 dig	Variant 2
4	Unit groups	436	4 dig	Occupations	436	4 dig	Variant 1

In both coding processes Cascot is used for automatic and manual coding of occupation. The program uses a classification file which includes an index or search list of occupations and rules have been drawn up to enhance the coding quality, for example by entering abbreviations or synonyms. During the coding of occupation, the Cascot program provides all codings with a score. This score is a measure of the probability that the assigned code is correct. The higher the score, the better the quality. In the coding process the score is an important control variable that is applied differently for each of the two processes, see sections 3.1 and 4.1. Sections 3.4 and 4.4 discuss the quality of the output that was ultimately delivered. In chapter 5 background information is given on the way Cascot was adjusted to make it suitable for measuring occupations of the Dutch labour market.

3. Process flow variant 1

3.1 Coding steps

Process variant 1 is split into 4 steps. The Cascot program is used in the first two steps during automatic coding and in the 4th step during manual coding. In Cascot, a classification file is used that contains an index or search list with job titles as well as a set of search rules that help to assign the correct code to the text. In this process variant the classification file used during manual coding differs from the one used in the first two steps during automatic coding.

3.1.1 Step 1: input variable occupation

In the first step of the coding process the respondents' answers to the question on occupation are coded automatically. The list of job titles per ISCO unit group that can be consulted in the ISCO 2008 manual[7] of the International Labor Organization (ILO) is the starting point in defining job titles that can be coded without additional information on tasks or further supporting information. This concerns job titles where the main tasks that are usually performed are also characteristic of the set of tasks that define a specific ISCO unit group. For example, the job title 'sociologist' can be provided with an ISCO code without considering a description of the respondent on the main tasks, in contrast to the job title 'scientist' that is too vague to code directly without additional information.

In this step records are provided with a valid ISCO 2008 unit group code. This is the lowest ISCO 2008 aggregation level (4 digit code). Not-further-defined-codes[8] (nfd-codes) used for descriptions of jobs that can not further be defined are not used in this step. For certain job titles derivation codes[9] are assigned.

The records that have been coded with a score of 40 and over and without a derivation code end up in the output, the rest continues to the next step. Around 60% of all the answers are coded in this step, see table 2.

3.1.2 Step 2: input variables for job title and tasks

In the second step answers to the questions on job title and main tasks are combined into a text field that constitutes the input for automatic coding. For example, the answer 'mechanic', which is too general to be coded in the 1st coding step, may be coded successfully by combining this job title with a description of tasks such as 'repairing cars'.

---

[7] http://www.ilo.org/public/english/bureau/stat/isco/isco08/

[8] Not further defined-codes are made up of a higher aggregation level code completed with one or more trailing zeros so as to make a 4 digit code. They are used to code responses that are too vague or broad to code at a more detailed level. The response can be assigned to an code for what is effectively a artificial unit group by adding trailing zeros to the aggregation level of the classification that cannot be further defined.

[9] A derivation-code is an auxiliary 5-digit code that is assigned to a specific selection of occupations through the Cascot default coding rules. For this reason, the derivation codes are also included in the search index and in the structure of the Cascot classification file. The derivation codes serve as input for coding step 3.

The records that have been coded with a valid ISCO-code along with a score of 70 and over end up in the output. Around 2% of all the answers to be coded are coded in this step, see table 2.

The portion that has been assigned a derivation code will pass to step 3. If a derivation code was assigned in both steps, the derivation code of the first step will be given priority.

3.1.3 Step 3: derivation using managerial tasks and economic activity

In the third step a derivation diagram is followed by means of which answers containing job titles that have in the first two steps been provided with a derivation code can be assigned ISCO-codes under specific conditions of auxiliary variables. In principle, this applies to job titles that are too briefly described to enable assigning an ISCO code on the basis thereof, for example director, entrepreneur, account manager, sales manager, or mechanic. The economic activity and answers to the questions on managerial tasks are then used for allocating an ISCO code.

In this step records that have in either step 1 or step 2 been assigned ISCO-codes of major group 1 'Managers' are also forwarded to manual coding if it is known that the person has no managerial tasks.

Records of which the auxiliary variables meet the specified conditions are assigned 4-digit ISCO codes and arrive in the output, the rest continues to the next step, manual coding.

Around 12% of all answers are coded in this step, see table 2.

3.1.4 Step 4: manual coding

In the fourth step all records which could not be provided with a code in the preceding steps are coded manually, because either the score is lower than 40 in step 1, or lower than 70 in step 2, or the derivation rules in step 3 do not apply.

In addition to the variables already used, coders use information on the number of persons supervised and the educational level attended. During manual coding, it is permitted to code at a higher aggregation level of the ISCO classification. In this case not-further-defined-codes are used.

Around 28% is coded manually in this step.

Table 2. Distribution of the coding steps, LFS 201612 to LFS 201707, unweighted, 40 thousand records.

Coding step	relative share to total
	%
Automatic	
1-job title	58
2-job title & tasks	2
3-auxiliary variables	12
Total automatic	72
Manual	
4-manual codng	28
Total	100

The classification file used in this step differs from the one used in the first 2 steps of the coding process.

3.2 Input

Process variant 1 can only be used if the occupation is observed via the CBS standard interrogation method[10]. The following variables from this questionnaire are used during coding:
• During automatic coding: job title, main tasks, managerial tasks, exclusively managerial, partly managerial, human resource management, strategic policy and the economic activity code.
• During manual coding, the following additional variables are used: educational attainment, number of persons supervised, number of staff of self-employed persons.

The CBS questionnaire on occupation is designed to measure occupations according to ISCO 2008. The questionnaire reflects the ILO recommendations[11] with regard to the observation of occupation in surveys and contains open questions on the occupation and the main tasks. In addition, Statistics Netherlands asks a number of additional questions on the type of managerial tasks to enhance the coding of the managers.

The first question deals with the respondent's occupation.

Occupation / *Beroep*

What is ($ A: your $ B: his $ C: her) occupation or what job do(es) ($ A: you $ B: he $ C: she) perform?

>> INT .: Try to be as specific as possible in the description, for example by including a specialization or level

Therefore not:	But rather:
Manager	Manager automation, Manager care, Financial Manager
Nurse	Psychiatric nurse, Nurse level 4, Nurse at the emergency room
Mechanic	Car mechanic, Electric engineer, Machine mechanic <<

The following questions are closed questions on managerial tasks.

ManagerialTasks / *Leiding*
Do(es) ($ A: you $ B: he $ C: she) you supervise ($ A: your $ B: are $ C: its) any employees? [TyesNo]

N_Supervised / *NLeidw*
How many persons?
1. 1 - 4 [N_1tm4]
2. 5 - 9 [N_5tm9]
3. 10 - 19 [N_10tm19]
4. 20 - 49 [N_20tm49]

[10] CBS standard interrogation method is available on the website of Statistics Netherlands at methoden/onderzoek/aanvullende onderzoeksbeschrijvingen

[11] http://www.ilo.org/public/english/bureau/stat/isco/isco08/

5. 50 - 99 [N_50tm99]
6. 100 or more [N100more]

In the questionnaire section on economic activity self-employed are asked whether they employ staff and if so how many staff members they employ. Then they go through the same questions as the employees.

ExclusivelyManagerial / *Uitsleid*
Do(es) ($ A: you $ B: he $ C: she) merely have managerial tasks or do(es) ($ A: you $ B: he $ C: she) also perform the same tasks as the staff/employees ($ A: you $ B: he $ C: she) supervise(s) ?
1. Exclusively managerial [Exclusively]
2. In addition to managerial tasks the same tasks as staff/employees [SameTasks]

PartlyManagerial / *Deelleid*
What makes the greater part of ($ A: your $ B: his $ C: her) tasks?
1. Management [Leadership]
2. Other tasks [OtherTasks]

Only if the greater part of the work consists of supervising the question on decision-making power follows.

HumanResources / *Persbeleid*
Do(es) ($ A: you $ B: he $ C: she) have the authority to make decisions in human resources such as hiring staff or giving a pay rise?
[TyesNo]

StrategicPolicy / *Stratbeleid*
Do(es) ($ A: you $ B: he $ C: she) have the authority to make decisions regarding the financial or strategic policy of the organization, such as the budget or the multi-year plan?
[TyesNo]

Finally, respondents are asked a question on the main tasks.

MainTasks / *VoornWzh*
What are ($ 1: besides management) the main tasks that ($ A: you $ B: he $ C: she) perform(s)?

>> INT .: Try to be as specific as possible in the description.

Therefore not:	But rather:
Advising	Advising private individuals on mortgages, advising students for further education, providing companies with legal advice
Administrative tasks	Administrative bookkeeping, keeping student administration, invoicing, data entry

Caring Care for children, providing home care for the elderly, care for the disabled

<<

The open questions on occupation and tasks have deliberately been designed such to encourage respondents to describe their occupation and tasks as accurately as possible in order to prevent that occupations will not be coded due to a lack of information.

In addition to the variables that are observed in the occupation section, the coding process also uses information on the economic activity (SBI), size of company and the level of the highest education attended. Economic activity is used during automatic coding in step 3 of the coding process, the company size and level of the highest education attended is only used in the 4th coding step during manual coding.

Variable for education level:

Der_HgstLevAtt / *Afl_HgstNivGev*
* Derivation of highest level of education attended

1. Lbo, vso (lts, leao, vbo, huishoudschool, ambachtschool) [LBO]
2. Vmbo, lwoo (including theoretische leerweg) [VMBO]
3. Mavo (ulo, mulo) [Mavo]
4. Havo (mms) [Havo]
5. VWO, gymnasium, atheneum (hbs, lyceum) [VWO]
6. Mbo (mts, meao, middenstandsdiploma, pdb, mba) [MBO]
7. HBO (hts, heao, kweekschool, associate degree) [HBO]
8. University education, including postgraduate courses and doctoral research, [Univ]
9. Other (company)training or course [Course]

Variable for company size in business section (for self-employed):

CompanySize[i] / *BedrOmvz*
($ 1: How many people ($ 1: $ A: do you $ B: does he $ C: does she) approximately employ?
$ 2: ($ A: Including yourself $ B: including himself $ C: including herself), how many people ($ A: does your partner $ B: does his partner $ C: does her partner) approximately employ ?
$ 3: ($ A: Including yourself $ B: including himself $ C: including herself), how many people do ($ A: your in-laws $ B: his in-laws $ C: her in-laws) approximately employ?

1. 1 [N_1]
2. 2 - 4 [N_2tm4]
3. 5 - 9 [N_5tm9]
4. 10 - 19 [N_10tm19]
5. 20 - 49 [N_20tm49]
6. 50 - 99 [N_50tm99]
7. 100 or more [N100more]

The variable for company size goes into some more detail than the answer categories of N_Supervised, therefore for manual coding the categories 1.1 and 2. 2-4 are merged into a category supervising 1-4 people.

3.3 Output

Process variant 1 by default generates ISCO unit group codes, the lowest aggregation level of the ISCO-2008 classification. In this coding process, coding at a higher aggregation level of the ISCO 2008 or with 'occupation unknown' is only used during manual coding in case the input contains too little information to code in more detail. The not-further-defined-codes that are produced in these cases can be linked to the major, sub-major, and minor groups of the ISCO 2008, or the occupation classes, segments, groups of the BRC 2014[12].

Table 3 shows that after completion of the coding steps 96.3% was coded on by an ISCO 4-digit code, and only 3.7% was coded at a higher aggregation level of ISCO 2008 using a not-further-defined-codes. The proportion of occupation coded unknown is low, just 1.5%.

Table 3. Share coded using nfd-codes at a higher aggregation ISCO 2008 level or unit group codes or occupation unknown after completion of the coding steps of process variant 1, LFS 2016 before weighting, 92 thousand records

Aggregation level ISCO 2008	Relative share to total %
nfd-codes major groups	1,5
nfd-codes sub-major groups	1,3
nfd-codes minor groups	0,9
Unit groups	96,3
Unknown	1,5
Totaal	100

3.4 Quality

The quality of process variant 1 is monitored by means of an annual random check of a few thousand records to verify whether the 4-digit codes are correct and in accordance with the definition of the unit groups of the ISCO 2008 manual and the available information at the moment of coding.

Compared to the overall total of codings, the share of incorrect 4-digit codes must not exceed 5%, per output category reported in regular statistics on Statline (3rd aggregation level of ISCO or BRC), the proportion of incorrect coding must not exceed 10%. Errors found may lead to modification of the automatic coding process.

Results are annually discussed with various stakeholders in the coding process: process coordinators and programmers, project managers of the various research projects and coding experts of occupation. Exchange of experiences may result in extra entries in the index or additional rules of Cascot or modifications in the coding process.

[12]Available on the website of Statistics Netherlands at methoden/classificaties/beroepenclassificaties/codelijsten en beroepenindex.xls

4. Process flow variant 2

4.1 Coding steps

Process variant 2 is split into 3 steps, the 3rd coding step 'manual coding' being optional. In this process coding takes place in Cascot using a classification file. This classification file is identical to the classification file used in process variant 1 during manual coding.

4.1.1 Step 1: input variable occupation

In the first step of the coding process, the answers to the question on occupation are coded automatically. In this step all records are provided with a valid ISCO 2008 unit group code.

All records continue to the next step.

4.1.2 Step 2: input variables occupation and tasks

In the second step answers to the questions on occupation and main tasks are combined into a text field that serves as input for automatic coding in this step. In this step all records are also provided with a valid ISCO 2008 unit group code.

After the second coding step the scores of the records coded in steps 1 and steps 2 are cross-matched. The coding with the highest score ends up in the output. In case of equal scores the code from step 1 is accepted. Records coded occupation unknown in the first step and with a valid ISCO code in the 2nd step only arrive in the output if the score in the second step is 40 or higher. The remainder is assigned the ISCO code for occupation unknown.

The dataset coded in these first two steps is included in the output database if there is no need to improve the quality and to use manual coding to this end.

Table 4 shows that in case of fully automatically coding 83% is coded on the basis of the answers to the questions on occupation and 17% on the basis of the combination of the answers to the questions on job title and tasks. The distribution of coding of occupation on the basis of the combination of job title and tasks deviates considerably from process variant 1 (table 2), because in that variant only automatically 4-digit coded records of sufficient quality arrive in the output. The delineation of the fraction of the records automatically coded in process variant 2 of sufficient quality will be made after completion of the first 2 coding steps, see also section 4.1.3.

Table 4 relative distribution of coding steps after fully automatic coding, LFS 2016 unweighted, 92 thousand records

Coding step	Relative share to total %
Automatic	
1-job title	83
2-job title & tasks	17
Total	100

4.1.3 Step 3: optional manual coding

If the client has opted for manual coding to obtain sufficient quality of all categories within the desired aggregation level of the classification of occupation, a selection is made of the categories of insufficient quality after fully automatic coding (see section 4.4). Of those categories, the fraction in the first two steps coded with a score lower than 40[13] will then be coded manually. In making this selection, the client's need for detail is taken into account to avoid any unnecessary manual coding.

During manual coding coders have access to the variables that are also used during manual coding in process variant 1, i.e. economic activity (SBI), managerial tasks, number of subordinates and educational level.

The records that have been coded with 'occupation unknown' after the first two coding steps will not again be coded manually in process variant 2. The reason for this is that the proportion of occupation unknown is only 4% compared to the total (see table 5). In comparison with process variant 1 this share is higher (see table 3), but the additional records coded unknown in variant 2 when coded according to process variant 1 occur in all occupational classes or major groups and are not concentrated in just a few thereof (see appendix 4). This also applies to the distributions of the more detailed aggregation levels of ISCO 2008 or BRC 2014. Omission of manual coding of the unknown will cause the share of unknown in process variant 2 to be higher, but this will not affect the quality of the remainder with valid ISCO codes.

4.2 Input

In process variant 2 just the variables occupation and tasks are used during automatic coding, other auxiliary information is used during manual coding if there is a need to improve the quality. If observed via the standard CBS question, the following variables are included:

• During automatic coding: Occupation, Tasks (see section 3.2).
• During manual coding, if available, the following are also used: Leiding, UitsLeid, DeelLeid, PersBeleid, StratBeleid, de SBI-code, Afl_HgstNivGev, N_LeidW, BedrOmvZ (see section 3.2 for an explanation of the variables).

If an external party has developed its own questions on occupation, it is under certain conditions possible to have the collected answers coded by Statistics Netherlands. It should concern open questions on occupation and by preference combined with a question on the main tasks. Other relevant occupation variables can only be used if observed in the same way as by Statistics Netherlands, or if these variables can be converted that way.

4.3 Output

The output that is produced depends on the client's wishes with regard to the aggregation level of ISCO 2008 or BRC 2014 and quality. If fully automatic coding has been chosen, the output will be delivered on 4 digits (including not-further-defined-codes) of ISCO 2008. Only a minor part, less than

[13] In this step, if the standard questionnaire of Statistics Netherlands for occupation is used, the variables on management will also be used to route to manual coding the records that in steps 1 and 2 were assigned ISCO major group code 1 Managers for persons without managerial tasks. This procedure is applied regardless of the score.

5% is coded with nfd-codes, see table 5. In addition, the score for each code is provided, which is a measure of the degree of the probability that the coding is correct. With the aid of mappings these codes can be converted to the higher aggregation levels of BRC 2014 and to ISCO 2008. Appendices 2 and 3 provide insight into the quality that is then achieved per category.

Table 5. Share of fully automatically coded using ISCO 2008 codes at a higher aggregation level by means of nfd-codes, unit group codes, or occupation unknown after completion of steps 1 and 2 of process variant 2, LFS 2016 before weighting, 92 thousand records

Aggregation level ISCO 2008	Relative share to total
	%
Nfd-codes major groups	2,4
Nfd-codes sub-major groups	1,3
Nfd-codes minor groups	0,3
Unit groups	92,1
Unknown	3,9
Total	100

If the client has opted for manual coding to obtain sufficient quality of all categories within the desired aggregation level of the classification, the requested level of aggregation will be provided or can be obtained via the SSB-method[14].

4.4 Quality after manual coding

The quality achieved after the first two automatic coding steps of process variant 2 is made measurable by using the LFS annual data files coded via process variant 1 as reference material. The size of this dataset is adequate for determining which part has yet to be coded manually to enhance the quality for all aggregation levels and underlying groups of BRC 2014 and ISCO 2008.

From a regression analysis information was obtained on the relationship between the share of unequal codes between the two processes and the share of incorrect codes from a manually checked sample of 500 records, see appendix 1. This showed that in order to meet the quality criterion of a maximum of 10% incorrect codes per category from the 1st to 3rd aggregation level of BRC 2014 or ISCO 2008 the percentage of unequal codes must not exceed 10%.

Table 6 illustrates, based on an annual LFS data file, that a decreasing score[15] goes hand in hand with an increasing share of unequal codes between the two processes (fully automatically coded through process variant 2 compared with coded through process variant 1). It also appears that the proportion of unequal codes is higher when the codes are compared in a more detailed level of aggregation. Overall, for each level of aggregation, the quality using BRC 2014 is slightly better than ISCO 2008. For example, when aggregated to the occupational classes of BRC2014, after full automatic coding

[14] In the System of Social Statistical datasets (SSB) ISCO 4-digit codes (including nfd-codes) are stored, through reference books the aggregation level with sufficient quality can be linked.

[15] During the coding of occupation the Cascot program provides all codings with a score. The score is a measure of the probability that the assigned code is correct. The higher the score, the better the quality.

15% was coded into a different occupational class than in process variant 1, but aggregated to major groups of ISCO 2008 17% were assigned different major group codes.

This table also shows that if a threshold score of at least 40 is applied, the share of unequal codes in the LFS in the 1st to 3rd aggregation level of the BRC 2014 or ISCO 2008 is 11% or less. On the basis of this fact this threshold value, a score of at least 40, is applied in the process to demarcate the portion that remains to be coded manually to improve the quality of the total of all types within the desired aggregation level. The portion scoring less than 40 is coded manually when within a certain category the share of an unequal codes exceeds 10%.

Table 6 Distribution of automatic coding over the score classes, and for each score category the share of unequal codes (fully automatically coded through process variant 2 compared to coded via process variant 1 by aggregation level of ISCO 2008 and BRC 2014, LFS 2016 unweighted.

			BRC 2014			ISCO 2008				
			Class (2 dig)	Segment (3 dig)	Group (4 dig)	Major (1 dig)	Sub-major (2 dig)	Minor (3 dig)	Unit (4 dig)	Level (1 dig)
Score class	# by score class	% to total	% unequal codes by score class			% unequal codes by score class				
90-100	26 034	28	1	1	1	1	1	1	2	1
80-89	7 467	8	2	2	3	2	2	3	4	2
70-79	8 479	9	3	4	4	3	4	5	5	3
60-69	9 566	10	6	9	11	8	10	12	16	7
50-59	8 695	9	13	19	23	16	20	24	28	13
40-49	11 508	12	18	26	30	21	26	32	40	17
30-39	12 699	14	36	55	64	39	58	67	70	34
20-29	3 861	4	57	73	79	59	74	79	84	47
10-19	187	0	66	82	87	70	82	87	91	59
1-9	2	0	50	100	100	50	100	100	100	50
0 (occupation unknown)	3 577	4	76	76	76	81	81	81	81	81
Total	92 075	100	15	20	23	17	21	24	27	14
40 or more	71 749	78	6	9	10	7	9	11	13	6
1 to 40	16 749	18	41	59	67	44	62	70	74	37
0 (occupation unknown)	3 577	4	76	76	76	81	81	81	81	81
Totaal	92 075	100	15	20	23	17	21	24	27	14

The output quality obtained via process variant 2 after using manual coding was further investigated through a pilot carried out in 2017 in two surveys, the ICT survey (sample 3221 respondents) and AKO (sample 1318 respondents). Table 7 shows the distribution of the codings over the coding steps for the two studies. The required aggregation level of the classification of occupation to be produced with sufficient quality is not the same for the two surveys. For the ICT survey it is the second aggregation level, for AKO it is the third level of aggregation. In the ICT research 17% and in AKO 20% are still coded manually. For comparison, in process variant 1 28% is coded manually.

Table 7 Distribution of the coding steps in case of manual coding. For ICT, output requirement was the second level of aggregation, for AKO the output requirement was the third aggregation level. Unweighted data ICT more than 3200 records and AKO more than 1200 records.

Coding step	2nd aggregation level ICT %	3rd aggregation level AKO %
Automatic		
1-job title	71	63
2-job title & tasks	12	16
3-manual coding	17	20
Total	100	100

For the ICT study the quality after manual coding was assessed by occupational segment (2nd aggregation level) of BRC 2014[16]. Compared to the overall total of codings, only 3% was assigned an incorrect code. In 33 of the 41 occupational segments, the proportion of incorrect codings is less than 10%. In 8 segments[17] the share of incorrect codes varies between 10 and 15%. This shows that in studies based on a much smaller sample size than the LFS, fluctuations around the target value of a maximum of 10% should be taken into account, but that the overall quality is good.

The quality of automatic coding after passing through the first two coding steps of process variant 2 is annually monitored on the basis of an updated LFS annual data file by checking whether changes occur in the share of unequal codes and whether it is necessary to make adjustments to the delineation of the proportion to be coded manually.

4.5 Quality after fully automatic coding

In appendices 2 and 3 the tables are presented that provide insight into the quality per category in the various aggregation levels, if full automatic coding is chosen. For this purpose, LFS reporting year 2016 was used, a sample of 92 thousand records. Table 8 below shows a fragment which illustrates how this should be interpreted.

The first two columns show the relative size of the group for each of the two coding processes per category of the 1st, 2nd and 3rd aggregation levels of BRC 2014 (appendix 2) and ISCO 2008 (appendix 3). The first column shows the proportion relative to the total when data is fully automatically coded, the second shows the proportion in case of coding according to process variant 1. This provides insight into the extent of the change in relative size of specific groups within the higher aggregation levels after fully automatic coding compared with process variant 1.

The third column shows for each fully automatically coded category the calculated share that was assigned a different code in process variant 1. Manual coding can be applied to improve the quality within specific aggregation levels for records with a score less than 40 within categories with a share of unequal codes exceeding 10%, see section 4.4.

[16] Internal note, report pilot ICT AKO nieuw typeerproces beroep_v08012018. doc

[17] This concerns 021- Authors and artists; 032- Sales representatives and buyers; 061- Government officials and managers;075- Food processing occupations and crafts n.e.c.; 076- Electricians and electronics mechanics; 078- Labourers construction and industry; 082- Associate professionals ICT en 092- Labourers agricultural, Internal note Resultaten pilot ICT en AKO, typeerproces beroep typeren hoger aggregatieniveau, 08-01-2018

For each category of BRC 2014 or ISCO 2008 table 8 also gives an indication of the expected quality in case of fully automatic coded data using the process variant 2. This indication is based on an estimate of the share of incorrect codes according to model C (see appendix 1) of the regression analysis in which the relation between the shares of unequal and incorrect codings was calculated. The estimated share of correct coding is split into 5 categories ranging from 'good' with an estimated share of incorrect from 0-9% to 'very bad' with an estimated share of more than 50% incorrect.

Table 8 (fragment of appendix 2) The distribution of the coding of the occupational classes, segments and groups of BRC 2014 according to process variant 1 and fully automatically coded according to process variant 2, the proportion per fully automatically coded category with a different code in process variant 1 and a quality indication, LFS 2016 unweighted total 92 thousand records.

		Process variant 2 (fully automatic)	Process variant 1	Proportion of unequal codes per fully automatically	Quality indication for fully automatic coding on the basis of estimated proportion incorrect				
		Proportion to total	Proportion to total	coded category	Good (0-9% in-correct)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% in-correct)
BRC2014, Occupational class, segment, group		%	%	%					
01	Pedagogical occupations	7,16	7,09	9,6		x			
011	Teachers	4,87	4,94	8,1		x			
0111	University and higher education teachers	0,69	0,60	33,7				x	
0112	Vocational education teachers	0,34	0,40	38,5				x	
0113	Secondary education teachers	1,11	1,26	18,0			x		
>>>for remainder of this table, see appendix 2<<<									
Total		100	100						

If fully automatic coded data is selected for analysis and publication purposes, it is advisable to take into account the quality before making statements with respect to individual categories of ISCO 2008 or BRC 2014 and to include a reference to documentation that can be consulted for the expected quality.

Appendix 4 provides insight into how fully automatic coding is distributed over the major groups of ISCO 2008 or the occupational classes of BRC 2014 after coding using process variant 1. This illustrates where records with unequal codes according to process variant 1 will be classified. For example, 90.4% of records coded in professional class 1 'Pedagogic occupations' after full automatic coding are also coded in the same occupational class using process variant 1. The remaining records are spread over all other occupational classes, but mainly in occupational class 10 ' Care and welfare 'and 4 ' Business and administrative occupations'. If needed, these tables can also be supplied for the 2nd or 3rd aggregation levels of ISCO 2008 or BRC 2014.

5. Implementing Cascot

The two processes that were developed to measure occupations according to ISCO 2008 use Cascot[18] as coding tool. Cascot uses a classification file that describes the structure, the index and rules separately and can be exported as separate files. The structure consists of a list of categories, each with a unique code and title according to the classification. The index is a collection of text descriptions, each associated with a specific category within a classification. The classification may have a number of rules associated with it, rules are however optional. They are designed to standardize coding procedures and are often derived from coding practices.

Cascot is designed to assign a code to a piece of text. It performs a complicated analysis of the words in the text, compares them to the words in the index files, and provides a list of recommendations. When compiling this list of recommendations Cascot also calculates a score from 0 to 100 which approximates the probability that the code recommended for a specific piece of input text is correct.

During implementation of Cascot in our coding process we separated measurement of quality and of performance and aimed to achieve at least 60% coded automatically with sufficient quality. The scoring proved to be a very helpful tool in this process. It can be used as an indication of the quality of the coding, and to separate the portion with too low quality, which needs to be manually coded, from the portion with sufficient quality.

In the following paragraphs it is explained how we optimized the index, and the rules we found most useful for coding occupations.

5.1 Development of the index

The project of adjusting Cascot for coding job titles collected in the Dutch LFS using a new design with interviewing via the internet, phone and face-to-face started in 2010. At the time the revised ISCO 2008 had to be implemented in 2011, the system needed to be adjusted to measure occupation in a back-office coding facility and according to another classification than the one we had used so far.

Until then the system that was used at Statistics Netherlands for coding occupation combined computer assisted interactive coding during the interview and automatic coding in batch of the remaining portion that could not be coded during the interview. The main portion was coded by the interviewers during the interview. This system used as a search file a huge database that contained manually coded open text answers to the question on occupation of several years of LFS-data (1985 and onwards). The codes that were assigned during the interview were so-called provisional codes. These codes had to be converted afterwards to the classification codes of the ISCO 2008 (or the national classification of occupation, SBC 1992) on the basis of additional conditions of several variables. The characteristics of the system made it difficult to use its content directly for implementation in Cascot. Therefore we started by investigating the suitability of other already existing lists of job titles with ISCO 2008 codes assigned to them as an index file for use in Cascot.

We compared the coding results using three different index files. One index file made use of the list of Dutch occupational titles that was developed in the Euroccupations project. This is an EU-funded project in which a detailed occupation database was developed for comparative socio-economic

[18] https://warwick.ac.uk/fac/soc/ier/software/cascot/

research in the EU. It contained the 1500-2000 most frequent occupations in the 8 largest EU-countries. The other two index files were based on our own national classifications of occupation (SBC 2010 and SBC 1992), one contained the official list with 19,000 unique occupational titles including detailing and synonyms. In the other index file we supplemented this list with job titles that were used in the coding process itself during manual or automatic coding. This most extended list contains 30,000 job titles in total. All job titles are unique, although in some cases the only difference is the order in which words are combined, or an addition of level of education is applicable. The difference between the three index files is illustrated by means of the job title 'civil engineer roads and waterworks' coded in ISCO unit group 2142 'civil engineers'.

In index 1 ISCO-code 2142 contains 4 entries, with one for the engineer civil engineering roads and waterworks 'Ingenieur weg- en waterbouwkunde (wo)'.

In index 2 this ISCO-code contains 63 entries in total with 8 job titles related to civil engineering roads and waterworks 'ingenieur weg- en waterbouwkunde (wo) differentiating between the materials worked on or with, or the kind of tasks performed (advising, researching, designing, constructing)

adviseur beton- en staalconstructies weg- en waterbouw
betonconstructeur weg- en waterbouw
constructeur staalbeton weg- en waterbouw
constructeur staalbouw weg- en waterbouw
constructeur weg- en waterbouw
ontwerper-constructeur weg- en waterbouw
staalconstructeur weg- en waterbouw
wetenschappelijk onderzoeker weg- en waterbouwkunde

In index 3 ISCO-code 2142 contains 118 job titles with 33 job titles related to civil engineering roads and waterworks, in this list more variants are included that relate to skill specialization, and sometimes includes information between brackets to help the coding expert during manual coding.

adviseur beton- en staalconstructies weg- en waterbouw
assistent-projectleider weg- en waterbouw (incl ontwerp)
betonbouwkundige weg- en waterbouw
betonconstructeur weg- en waterbouw
betonconstructeur-tekenaar weg- en waterbouw
civiel ingenieur weg- en waterbouw (ontwerpen-construeren)
constructeur staal-, betonconstructies weg- en waterbouw
constructeur staalbeton weg- en waterbouw
constructeur staalbouw weg- en waterbouw
constructeur staalconstructies weg- en waterbouw
constructeur staalconstructiewerk weg- en waterbouw
constructeur weg- en waterbouw
groepsleider weg- en waterbouw (incl ontwerp)
ontwerper weg- en waterbouw
ontwerper weg- en waterbouw (ir)
ontwerper weg- en waterbouw (znd)
ontwerper-constructeur betonconstructies weg- en waterbouw
ontwerper-constructeur staal-, betonconstr weg- en waterbouw (ir)
ontwerper-constructeur staal-, betonconstr weg- en waterbouw (znd)
ontwerper-constructeur weg- en waterbouw
ontwerper-constructeur weg- en waterbouw (ir)
ontwerper-constructeur weg- en waterbouw znd
staalbetonconstructeur weg- en waterbouw
staalbouwkundig constructeur weg- en waterbouw
staticus (berekenen staal- en betonconstructies) weg- en waterbouw
weg- en waterbouwkundig constructeur znd

weg- en waterbouwkundig ingenieur (ontwerpen-construeren)
weg- en waterbouwkundig ontwerper (assistent projectleider)
weg- en waterbouwkundig ontwerper-constructeur
weg- en waterbouwkundige (niv h/m)
weg- en waterbouwkundige (wetens onderzoeker, schrijft wetens art)
wetens onderzoeker weg- en waterbouwkunde (schrijft wetens art)
wetenschappelijk onderzoeker weg- en waterbouwkunde

We tested the performance of automatic coding by using two different input files. We used the answers to open question on occupational titles collected during two years of the Dutch Labour Force Survey (2004 and 2005), and a selection of the thousand most frequently occurring job titles of these years. In total respondents used almost 50 thousand unique descriptions. The top 1000 most frequently occurring job titles represent approximately 50% of all respondents.

The quality of coding performance was measured by splitting up the coding results into 4 (non exclusive) groups with different quality [19]. One group covers records that are coded with a score of 100 and have an exact match with an index entry. Another group refers to records for which no match can be found and that are coded with ISCO-code unknown with a score of 0. The other two groups represent records with a score of 70 or higher, so with a relatively high degree of certainty that the chosen code is correct, and with a score of 40 and higher, so with a lower certainty.

When comparing the three index files the most extensive list that contains 30,000 job titles performs best. However, there is hardly any difference between the list with 19,000 job titles and the list with 30,000 job titles. In both input files the share of records with a score of 70 or higher is only increased by 1-2 % points when using the most extended list.

When comparing index file 3 with 30,000 index entries to index file 1 with 1600 job titles the share of records coded with a high score (70 or higher) increases from 35% to 50% when coding input file 1, and from 8% to 22% when coding input file 2. Given the fact that index file 3 contains almost 19 times more index entries than index file 1, one might expect a better performance than only an increase by approximately a factor 2.

Another striking finding was that using the index file 2 with 19,000 or index file 3 with 30,000 different job titles still only 50% of the most frequently occurring job titles (input file 1) can be coded with a high degree of certainty that the coding is correct, so with a scoring of 70 or higher. This share is only 20% when using the complete set of job titles collected in two years of LFS (input file 2).

See also table 6, where the results are shown of the classification file that is currently used, then 45% is coded with a score of 70 or higher.

Table 9 Distribution of automatic coding of all job titles collected (input 2) and the thousand most frequently occurring job titles (input 1) over the score classes using three different index files with an increasing number of job titles, LFS 2004 and 2005 unweighted.

	Input1: top 1000		Input2: LFS 2004, 2005	
Index file 1: 1600 job titles				
score 100	66	7%	299	1%
score 70 and higher	337	35%	3814	8%
score 40 and higher	642	66%	16814	34%
score 0	106	11%	5028	10%

[19] The relation between the scoring and quality of coding in the final version of the classification file is also illustrated in table 6.

Index file 2: 19 000 job titles				
score 100	81	8%	715	1%
score 70 and higher	473	49%	9903	20%
score 40 and higher	861	88%	29014	59%
score 0	30	3%	1669	3%
Index file 3: 30 000 job titles				
score 100	60	6%	593	1%
score 70 and higher	487	50%	10717	22%
score 40 and higher	882	90%	30161	61%
score 0	23	2%	1378	3%
total	975	100%	49522	100%

From these findings we drew the following conclusion relevant for further development of the classification file and implementation into the coding process:

1. There is a considerable mismatch between job titles listed in the classification of occupations and the way respondents describe their jobs. A more extended list of occupational titles does not necessarily lead to improvement of quality and performance of automatic coding.
2. To maximise the share of records coded automatically with high quality, the focus should be on developing an index that at least contains job titles that occur most frequently and that can be coded with an ISCO-unit group by using the wording/phrasing that occurs in the field. For further improvement we selected approximately 5000 job titles from the three tested index files. We included the Euroccupations list, added the titles with an exact match to answers of respondents relevant for coding the thousand most frequently occurring titles and supplemented them with detailing for answers that are often too vague to code to ISCO 2008 unit groups: researcher, advisor, engineer, account manager etc.
3. The rules can be used to provide solutions for most frequently occurring respondents' answers that are too vague to assign a 4-digit ISCO-unit group, e.g. manager, office clerk, technician.
4. It needs to be investigated what the optimal threshold value of the score should be to separate automatically coded records with sufficient quality from thas to be coded manually. Coders need to be trained to know how to code job descriptions with help of relevant other variables according to the ISCO 2008. To improve the quality and performance of automatic and manual coding, the results need to be checked and updated on a regular basis.

5.2 Developing the rules

In addition to the index the rules were used to improve performance and quality. In the following paragraph the rules that we found most useful are explained. During coding Cascot applies the rules in the following order: Abbreviations, Replacements, Conclusions, Alternatives, Default coding, and below they are discussed in this order.

Abbrevations

In the abbreviations we specified the abbreviation variants that are used in description of jobs or tasks. Once specified, Cascot applies this rule to all other job titles in which these abbreviations are used.

Replacements

In the replacements we specified the synonyms and most frequently occurring spelling mistakes. The rules are used by Cascot in order, starting with the rule with the lowest number. This enables you to first specify the spelling mistakes and synonyms of one word, and in a separate rule the combination of more words that should be replaced by only one. The replacements can also be used to delete a piece of text and increase the score of the remaining text that is coded automatically.

Conclusions

In the conclusions we specified the job titles that Cascot should not provide with an ISCO 2008 code. Here Cascot will not conclude. In coding process variant 1 all job titles that cannot conclude are forwarded to step 2, where they are combined with the main tasks.

It is also possible to specify to code a piece of text as ambiguous. In that case Cascot codes with an ISCO-code belonging to the best matching index entry, but the scoring is by default put to 40. This principle is used in process variant 2 to avoid coding with high score of job titles that are often too vague to assign an ISCO-unit group code. After combining with main tasks a better matching index entry can be found with a higher score, depending of course on the way tasks are described.

Alternatives

We used the alternatives to lead certain job description to a job title in the index without replacing it. For example when a person mentions as a job description such as the word ‘magazijn’ meaning ‘warehouse’, the person is probably a ‘magazijn medewerker’, a ‘warehouse worker’, but you do not want to replace ‘magazijn’ because other combinations exist that should lead to other ISCO 2008 codes.

The alternatives can also be used to provide suggestions where to look, for example when ‘advisor’ is found, also entries with the term ‘consultant’ may be relevant. The index is optimized for the most frequently occurring job titles, so it includes for example only the job title ‘business consultant’ and not ‘business advisor’, or ‘student advisor’ and not ‘student consultant’. Since we do not have all alternatives in the index, and we do not use the same term in all cases, the alternatives are a convenient tool to find relevant index entries.

Default coding

These rules define a set of words and phrases that should be scored as though they were a different word or phrase specified in the index.

Default coding rules are applied differently in process variant 1 and variant 2.

In process variant 1 we use the default coding rules to assign a decision code to the job titles we want to lead into the third step of our coding process, so that is where we combine job titles with auxilliary information on economic activity and managerial tasks (see paragraph 3.1).

These decision codes have 5 digits starting with 99, they are added to the structure at the second level below major group 9 with 99 at the second level, and the 5-digit codes at the fifth lowest level. The decision codes also have to be added to the index to be able to assign them during automatic coding using the default coding rules.

So, for example in case a respondent describes his or her job title as ‘teacher’, many ISCO-codes are possible, however in the case he or she is working as a trainer at a fitness centre there is a good chance that ISCO code 3423 Fitness and Recreation Instructors and Programme Leaders is correct. The default

coding rules are used to assign the decision code, so in the case of the teacher code 99058 is assigned.. Then we use a decision table, that was programmed separately, to lead them to a specific ISCO 2008 code. Teacher with a NACE code for fitness center activities is probably a fitness instructor, for remaining cases where the ISCO-code remains uncertain because the economic activity is not decisive enough and main tasks should be used, the records are sent to step 4, manual coding.

So, in the decision table only job titles are coded automatically when it is considered that the NACE code or the managerial tasks supply sufficient relevant information to assign ISCO-codes.

In process variant 2 the default coding rules are not used for assigning decision rules, but to force Cascot to code a piece of text to a particular code and manipulate the scoring associated with it by adjusting the weight. For example if the job description contains (among others) the words ‘owner’ and ‘selling’ a default coding rule assigns this descprition to an index entry belonging to code 5221 ‘shop owners’ with a score below 40. The score indicates the low uncertainty that the chosen code is correct, and the record can be coded manually if needed.

5.3 Advantages and disadvantages of the coding system

The system that we developed for collecting and coding data on occupation has advantages and disadvantages. With the transisition to a CAWI first mixed-mode data collection, it was necessary to develop a system where the coding could be done without the help of an interviewer. Also the objective was to minimize the amount of manual coding while maintaining a certain quality level. A system was developed that it is suitable for measuring occupations in a design with the same questionnaire in the different interviewing modes: via the internet (CAWI), telephone (CATI) or face-to-face (CAPI). It turns out that the difference in quality of the coding between the modes is fairly small. Actually, the amount of manual coding is the smallest in CAWI: slightly less than 30 percent compared to slightly more than thirty percent in CAPI and CATI. This can de be due to the different composition of the response in each mode. Nevertheless, within seven out of ten major groups the amount of manual coding is the smallest in CAWI.

Another advantage is that we have reduced the interviewing time, and the costs needed to train the interviewers to learn how to assign codes. We now also avoid the risk of divergence in the assignment of codes because of interviewer interpretations. However a disadvantage of the system is that in the CAWI-mode no feedback can be given to the respondents when giving vague job titles. We tried to overcome this by formulating the questions in such a way that respondents are stimulated to give precise descriptions that help to a certain extent (see section 3.2). Another disadvantage of the system is that manual coding is still necessary to achieve sufficient quality. Coding experts need to be trained in the basic principles of ISCO 2008 to know how to deal with vague job titles and how to code job titles for which no exact matching suggestion is available in the index.

Overall, we are content using Cascot that enabled us to develop a coding process with a good balance between automatic coding percentages and coding quality. The need for manual coding can be adjusted depending on the quality demands and in this way costs can be reduced.

6. Abbreviations

AKO	Arbeidskrachten onderzoek Caribisch Nederland à Labour Force Survey on the Dutch Caribbean
BRC 2014	Beroepenindeling ROA CBS 2014 / A classification derived from the ISCO 2008 unit groups developed by Statistics Netherlands in close cooperation with Maastricht university education and labourmarket researchers and public employment service in which the 436 ISCO unit groups are regrouped into 114 occupational groups that better suit the Dutch Labour Market.
CAPI/CATI/CAWI	Computer assisted face-to-face-/telephone-/web-based interviewing
CASCOT	Computer Assisted Structured Coding Tool
Derivation-code	A diversion code is an auxiliary 5-digit code that is assigned to a specific selection of occupations through the Cascot default coding rules. For this reason, the derivation codes are also included in the search index and in the structure of the Cascot classification file. The derivation codes serve as input for coding step 3.
ICT-survey	Survey on the use of ICT by the population
ISCO 2008	International Standard Classification of Occupations 2008
LFS	Labour Force Survey
NACE	Classification of Economic Activities in the European Community
NEA	Nationale Enquete Arbeidsomstandigheden à National Survey on working onditions among employees
Nfd-code	Not further defined-codes are made up of a higher aggregation level code completed with one or more trailing zeros so as to make a 4 digit code. They are used to code responses that are too vague or broad to code at a more detailed level. The response can be assigned to an code for what is effectively a artificial unit group by adding trailing zeros to the aggregation level of the classification that cannot be further defined.
Score	During the coding of occupation the Cascot program provides all codings with a score. The score is a measure of the probability that the assigned code is correct. The higher the score, the better the quality.
SBI	The Dutch Standaard Bedrijfsindeling is based on the activity classification of the European Union, the NACE. The first four digits of the SBI are the four digits of the NACE.
ZEA	Zelfstandigen Enquete Arbeid Survey on working conditions among self-employed

Appendix 1. Measuring quality after fully automatic coding

In order to gain insight into the quality per category of the occupational classifications, the codes assigned in the LFS reporting year 2015 in process variant 1 are compared with the codes that have been assigned by fully automatic coding through process variant 2.

A random selection of 500 records with codes according to process variant 1 and by fully automatic coding was checked manually. Per record a check was performed whether the code assigned in the fully automatic coding process was correct and the code was corrected if incorrect.

To gain insight into the relationship between the percentages with codes that differ from the current process and the percentages with incorrect codes, the selection checked was split into different groups, see table 1a. A total of 59 different groups are distinguished. The relationship between the percentages unequal and incorrect is plotted in a scatter diagram, see figure 1. Both table 1a and figure 1 show a correlation between the proportions unequal and incorrect. For each group, the shares of unequal and incorrect codes are closer when the differences in coding between the two datasets are smaller. This is, for example, illustrated in table 1a by the groups coded with a score of 40 and more versus coded with a score less than 40.

As the figure shows a linear relationship, the precise relationship between these variables is calculated by means of a linear regression analysis applied to all groups (model A), the groups with a minimum size of 10 observations (model B), and groups with a minimum size of 25 observations (model C), see table 1b. The percentage incorrect is used as a dependent variable whereas the percentage unequal is used as an independent variable.

Table 1b shows that all models are significant, and that model C, i.e. exclusively consisting of groups of more than 25 observations is the most reliable. Based on the constant and coefficient of model C, the percentage unequal should not exceed 10% to yield a percentage of 10%, according to models A and B the percentage unequal should not exceed 9%.

Based on these findings the rule was applied in the analyses of the quality of the fully automatic coding process that if within a category the share with unequal codes of a maximum of 10%, the percentage with incorrect codes will not be higher than 10% and thus of sufficient quality.

Figure 1. The relationship between the percentages with unequal codes (x-axis) and the percentages with incorrect codes (y-axis) of the groups presented in table 1a, sample 500 records LFS 2015.

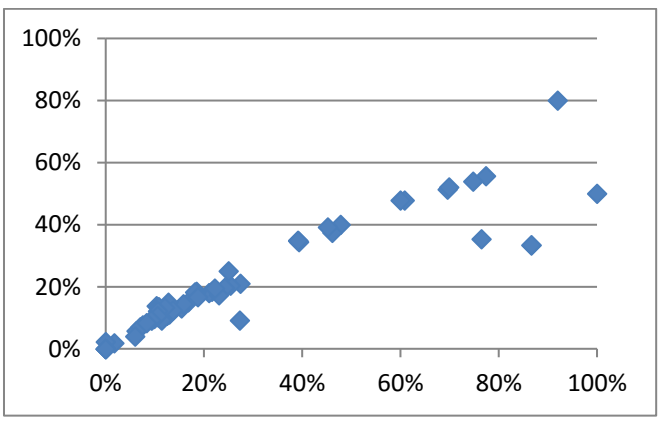


Table 1a. Fully automatically coding in the various aggregation levels of the ISCO and BRC and by score class with the share of unequal codes for process variant 1 per group and the share with incorrect codes, sample of 500 records from the LFS 2015. Orange shaded groups were omitted in model B , blue and orange shaded groups in model C.

	Number by group	% unequal	% incorrect
ISCO 2008			
Total			
Unit groups (4 digits)	500	27,4	21,0
Minor groups (3 digits)	500	25,4	20,4
Sub-major groups (2 digits)	500	21,6	18,4
Major groups (1 digit)	500	16,8	15,0
Level	500	15,4	13,0
Score 40 and over			
Unit groups (4 digits)	385	12,5	10,6
Minor groups (3 digits)	385	10,6	10,4
Sub-major groups (2 digits)	385	9,9	9,6
Major groups (1 digit)	385	7,5	7,5
Level	385	6,2	5,7
Score less than 40			
Unit groups (4 digits)	115	77,4	55,7
Minor groups (3 digits)	115	74,8	53,9
Sub-major groups (2 digits)	115	60,9	47,8
Major groups (1 digit)	115	47,8	40,0
Level	115	46,1	37,4
Major groups			
00~Armed forces occupations	1	0,0	0,0
01~Managers	23	39,1	34,8
02~Professionals	136	11,8	11,0
03~Technicians and associate professionals	65	18,5	18,5
04~Clerical support workers	44	11,4	9,1
05~Service and sales workers	116	10,3	13,8
06~Skilled agricultural, forestry and fishery workers	6	0,0	0,0
07~Craft and related trades workers	47	12,8	14,9
08~Plant and machine operators, and assemblers	11	27,3	9,1
09~Elementary occupations	36	22,2	19,4
Occuption unknown	15	86,7	33,3
BRC 2014			
Totaal			
Occupational group	500	24,2	19,6
Occupational segment	500	21,0	18,0
Occupational class	500	15,8	14,4
Score 40 and over			
Occupational group	385	10,6	10,1
Occupational segment	385	9,4	9,1
Occupational class	385	7,0	7,0
Score less than 40			
Occupational group	115	69,6	51,3
Occupational segment	115	60,0	47,8
Occupational class	115	45,2	39,1

	Number by group	% unequal	% incorrect
Occupational Classes			
01-Pedagogical occupations	37	10,8	13,5
02-Creative and linguistic occupations	16	25,0	25,0
03-Commercial occupations	73	13,7	12,3
04-Business economics and administrative occupations	83	12,0	13,3
05-Managers	23	39,1	34,8
06-Public administration, security and legal occupations	11	18,2	18,2
07-Technical occupations	66	10,6	12,1
08-ICT occupations	19	10,5	10,5
09-Agricultural occupations	7	0,0	0,0
10-Care and welfare occupations	76	9,2	9,2
11-Service occupations	48	18,8	16,7
12-Transport and logistics occupations	24	8,3	8,3
13-Not elsewhere classified	17	76,5	35,3
Total by scoreclass			
score100	10	0,0	0,0
score90-99	113	1,8	1,8
score80-89	45	0,0	2,2
score70-79	50	6,0	4,0
score60-69	54	13,0	11,1
score50-59	52	23,1	17,3
score40-49	61	39,3	34,4
score30-39	73	69,9	52,1
score20-29	25	92,0	80,0
score10-19	2	100,0	50,0
score0	15	86,7	33,3

Table 1b. Summary regression analysis of relationship between percentages unequal (x value) and percentages incorrect (y value): model A all groups, model B groups with minimum size of 10 observations, model C groups with minimum size of 25 observations.

			Unstandardized Coefficients		Standardized Coefficients			x-value if y=10%
	N		B	Std. Error	Beta	t	Sig.	
Model A	59	(Constant)	4,614	1,304		3,539	0,001	9,1
		unequal	0,589	0,034	0,918	17,477	0,000	
Model B	55	(Constant)	4,643	1,381		3,361	0,001	8,9
		unequal	0,605	0,037	0,915	16,546	0,000	
Model C	44	(Constant)	2,305	0,526		4,380	0,000	10,2
		unequal	0,751	0,015	0,991	49,298	0,000	

Appendix 2. Quality of fully automatic coding process BRC 2014

The distribution of the coding across the occupational classes, segments and groups of BRC 2014 according to process variant 1 and fully automatically coded according to process variant 2, the proportion per fully automatically coded category with a different code in process variant 1 and an indication of the quality, LFS 2016 unweighted total 92 thousand records.

		Process variant 2 (fully automatic) Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category[20]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect[21]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
BRC2014, Occupational class, segment, group		%	%	%					
01	**Pedagogical occupations**	**7,16**	**7,09**	**9,6**	**x**				
011	Teachers	4,87	4,94	8,1	x				
0111	University and higher education teachers	0,69	0,60	33,7			x		
0112	Vocational education teachers	0,34	0,40	38,5				x	
0113	Secondary education teachers	1,11	1,26	18,0		x			
0114	Primary school teachers	1,90	1,90	15,7		x			
0115	Educationalists and other teachers	0,82	0,78	31,2			x		
012	Sports instructors	0,51	0,56	19,4		x			
0121	Sports instructors	0,51	0,56	19,4		x			
013	Childcare workers and teaching assistants	1,79	1,58	21,8		x			
0131	Childcare workers and teaching assistants	1,79	1,58	21,8		x			
02	**Creative and linguistic occupations**	**2,30**	**2,07**	**22,7**		**x**			
021	Authors and artists	1,39	1,27	22,8		x			
0211	Librarians and curators	0,10	0,09	29,2			x		
0212	Authors and linguists	0,33	0,35	13,2		x			
0213	Journalists	0,30	0,28	21,6		x			
0214	Visual artists	0,12	0,13	2,8	x				
0215	Performing artists	0,55	0,43	35,9			x		

[20] Manual coding within a specific category is recommended when the share of unequal codes exceeds 10%, underlying categories with sufficient quality need not be coded manually. For example, occupational group 0214 Visual artists need not be coded, but the other occupational groups within segment 021 Authors and artists do to improve the quality of the entire segment.

[21] The proportion incorrect is calculated on the basis of a regression analysis into the relationship between the percentages of unequal codes and incorrect codes, as explained in appendix 1, model C.

		Process variant 2 (fully automatic) Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category[20]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect[21]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
BRC2014, Occupational class, segment, group		**%**	**%**	**%**					
022	Artistic and cultural specialists	0,91	0,79	25,4			x		
0221	Graphic designers and product designers	0,60	0,52	26,1			x		
0222	Photographers and interior designers	0,30	0,27	25,4			x		
03	**Commercial occupations**	**10,50**	**11,87**	**10,6**		**x**			
031	Advisors marketing, public relations and sales	1,42	1,65	26,4			x		
0311	Advisors marketing, public relations and sales	1,42	1,65	26,4			x		
032	Sales representatives and buyers	1,31	1,64	34,3			x		
0321	Representatives and buyers	1,31	1,64	34,3			x		
033	Salespersons	7,76	8,57	10,2	x				
0331	Retailers and team leaders retail	0,81	1,32	37,9				x	
0332	Retail sales staff	4,09	4,51	9,1	x				
0333	Cashiers	1,38	1,28	9,6	x				
0334	Call center employees outbound and other sellers	1,48	1,45	21,5		x			
04	**Business economics and administrative occupations**	**18,21**	**18,71**	**12,1**		**x**			
041	Professionals business management and administration	5,17	5,32	21,5		x			
0411	Accountants	1,26	1,06	25,4			x		
0412	Financial specialists and economists	0,89	0,83	29,6			x		
0413	Business consultants and management consultants	1,32	1,57	32,7			x		
0414	Policy advisors	0,82	0,72	22,7		x			
0415	Personnel and career development specialists	0,88	1,14	13,9		x			
042	Associate professionals in business management and administration	3,42	3,74	17,9		x			
0421	Bookkeepers	1,11	1,22	12,7		x			
0422	Business service providers	1,01	1,11	26,6			x		
0423	Executive secretaries	1,30	1,41	16,7		x			
043	Administrative staff	9,62	9,65	13,2		x			
0431	Administrative employees	3,35	3,55	21,3		x			
0432	Secretaries	0,72	0,63	21,1		x			
0433	Receptionists and telephone operators	2,16	1,93	21,9		x			
0434	Accounting staff	1,69	1,60	22,2		x			
0435	Transport planners and logistics staff	1,70	1,94	19,4		x			
05	**Managers**	**6,20**	**5,23**	**36,0**			**x**		
051	General directors	0,47	0,88	36,7			x		
0511	General directors	0,47	0,88	36,7			x		

BRC2014, Occupational class, segment, group		Process variant 2 (fully automatic) Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category[20]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect[21]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
		%	%	%					
052	Managers in the administrative and commercial area	2,21	1,48	45,4				x	
0521	Managers business and administrative services	1,14	0,91	33,6			x		
0522	Managers sales and marketing	1,07	0,57	59,1				x	
053	Managers production and specialized services	1,39	1,88	35,2			x		
0531	Managers production	0,25	0,55	36,9				x	
0532	Logistics managers	0,20	0,21	45,4				x	
0533	ICT managers	0,30	0,23	46,0				x	
0534	Managers health care institutions	0,20	0,33	23,4		x			
0535	Managers education	0,16	0,22	20,8		x			
0536	Managers specialized services	0,28	0,34	41,9				x	
054	Managers hospitality, retail and other services	0,26	0,68	35,9			x		
0541	Hotel and restaurant managers	0,12	0,19	27,2			x		
0542	Managers retail and wholesale trade	0,06	0,37	15,5		x			
0543	Managers commercial and personal services	0,07	0,12	69,2					x
055	Managers without specialisation	1,87	0,31	96,8					x
0551	Managers without specialisation	1,87	0,31	96,8					x
06	**Public administration, security and legal occupations**	**3,03**	**3,29**	**16,1**		**x**			
061	Government officials and managers	0,94	1,08	30,7			x		
0611	Government managers	0,24	0,41	38,7				x	
0612	Government officials	0,69	0,66	30,7			x		
062	Lawyers	0,67	0,76	7,2	x				
0621	Lawyers	0,67	0,76	7,2	x				
063	Security workers	1,42	1,45	15,5		x			
0631	Police inspectors	0,12	0,12	14,0		x			
0632	Police and fire department	0,41	0,41	15,2		x			
0633	Security personnel	0,64	0,60	17,0		x			
0634	Armed forces occupations	0,26	0,32	17,4		x			
07	**Technical occupations**	**12,27**	**12,85**	**13,8**		**x**			
071	Engineers and researchers in mathematical, physics and technical sciences	2,19	2,25	24,2			x		
0711	Biologists and natural scientists	0,41	0,33	38,0				x	
0712	Engineers (no electrical engineering))	1,22	1,45	23,5		x			
0713	Electrotechnical engineers	0,13	0,13	29,1			x		

		Process variant 2 (fully automatic) Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category[20]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect[21]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
BRC2014, Occupational class, segment, group		**%**	**%**	**%**					
0714	Architects	0,43	0,35	31,8			x		
072	Associate professionals in physics and technology	1,78	1,98	28,1			x		
0721	Technicians in construction and nature	1,01	1,08	27,9			x		
0722	Production leaders in industry and construction	0,51	0,65	28,4			x		
0723	Process operators	0,25	0,24	33,2			x		
073	Construction workers	2,53	2,61	12,6		x			
0731	Construction workers in structural work	0,61	0,65	24,2			x		
0732	Carpenters	0,79	0,80	9,6	x				
0733	Construction workers in finishing	0,36	0,37	14,3		x			
0734	Plumbers and pipe fitters	0,31	0,36	17,7		x			
0735	Painters and metal sprayers	0,46	0,43	12,2		x			
074	Metal workers, mechanics	1,91	1,92	15,6		x			
0741	Metal workers and construction workers	0,47	0,47	20,6		x			
0742	Welders and sheet metal workers	0,37	0,36	12,4		x			
0743	Car mechanics	0,61	0,61	16,1		x			
0744	Machine technicians	0,46	0,48	22,6		x			
075	Food processing occupations and crafts n.e.c.	1,36	1,33	25,3			x		
0751	Butchers	0,21	0,18	26,3			x		
0752	Bakers	0,23	0,22	11,4		x			
0753	Product inspectors	0,20	0,24	53,8				x	
0754	Furniture makers, tailors and upholsterers	0,40	0,40	17,8		x			
0755	Employees printing and crafts	0,32	0,29	28,5			x		
076	Electricians and electronics mechanics	0,79	0,87	24,1			x		
0761	Electricians and electronic mechanics	0,79	0,87	24,1			x		
077	Production machine operators and assembly workers	0,95	1,08	36,9			x		
0771	Production machine operators	0,74	0,79	39,5				x	
0772	Assembly workers	0,21	0,29	33,3			x		
078	Labourers construction and industry	0,78	0,81	21,2		x			
0781	Workers in construction and industry	0,78	0,81	21,2		x			
08	**ICT occupations**	**3,67**	**3,79**	**14,8**		**x**			
081	Professionals ICT	3,07	3,16	14,4		x			
0811	Software and application developers	2,28	2,39	16,7		x			
0812	Database and network specialists	0,79	0,77	12,4		x			

BRC2014, Occupational class, segment, group		Process variant 2 (fully automatic) Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category[20]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect[21]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
		%	%	%					
082	Associate professionals ICT	0,60	0,63	35,2			x		
0821	User support ICT	0,44	0,49	34,2			x		
0822	Radio and television technicians	0,17	0,14	39,1				x	
09	**Agricultural ocupations**	**1,88**	**2,11**	**9,0**	**x**				
091	Horticulturists, fieldcrop growers and livestock farmers	1,47	1,74	9,4	x				
0911	Crop growers and foresters	0,23	0,33	32,9			x		
0912	Gardeners, horticulturists and growers	0,79	0,86	12,9		x			
0913	Livestock farmers	0,46	0,55	15,1		x			
092	Labourers agricultural	0,40	0,37	24,8			x		
0921	Agricultural workers	0,40	0,37	24,8			x		
10	**Care and welfare occupations**	**13,39**	**13,12**	**9,7**	**x**				
101	Medical doctors, therapists and specialized nurses	3,55	3,33	16,9		x			
1011	Doctors	1,26	1,16	15,1		x			
1012	Specialized nurses	1,54	1,38	25,2			x		
1013	Physiotherapists	0,75	0,80	8,8	x				
102	Professionals in care and social work	1,49	1,63	15,4		x			
1021	Social workers	0,74	0,84	20,1		x			
1022	Psychologists and sociologists	0,75	0,80	11,7		x			
103	Healthcare associate professionals	2,69	2,83	15,6		x			
1031	Lab technicians	0,27	0,26	23,0		x			
1032	Pharmacy assistants	0,23	0,26	1,9	x				
1033	Nurses (secondary vocational ed.)	0,78	0,91	18,6		x			
1034	Medical practice assistants	0,80	0,84	3,8	x				
1035	Medical specialists	0,61	0,55	32,1			x		
104	Associate professionals in social work and residential homes	2,82	2,55	23,9			x		
1041	Social workers, group and housing supervisors	2,82	2,55	23,9			x		
105	Home-based and institutional personal care workers	2,83	2,78	11,9		x			
1051	Care workers	2,83	2,78	11,9		x			
11	**Service occupations**	**9,69**	**10,01**	**7,1**	**x**				
111	Personal service workers	5,50	5,79	8,8	x				
1111	Tourist guides	0,25	0,25	12,1		x			
1112	Cooks	0,68	0,78	4,8	x				
1113	Waiters and bar staff	2,91	3,11	5,7	x				

		Process variant 2 (fully automatic) Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category[20]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect[21]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
BRC2014, Occupational class, segment, group		**%**	**%**	**%**					
1114	Hairdressers and beauticians	0,72	0,78	3,0	x				
1115	Caretakers and cleaning team leaders	0,48	0,50	32,7			x		
1116	Other personal service providers	0,46	0,39	28,6			x		
112	Cleaners and kitchen helpers	4,18	4,22	8,0	x				
1121	Cleaners	3,08	3,07	8,9	x				
1122	Kitchen helpers	1,11	1,15	7,0	x				
12	**Transport en logistics occupations**	**7,59**	**7,80**	**5,6**	**x**				
121	Drivers vehicles and operators mobile machines	2,96	3,11	6,0	x				
1211	Deck officers and pilots	0,26	0,26	25,2			x		
1212	Car, taxi and van drivers	0,70	0,70	12,6		x			
1213	Bus drivers and tram drivers	0,25	0,22	16,7		x			
1214	Truck drivers	1,13	1,19	10,9		x			
1215	Mobile machine operators	0,62	0,73	6,3	x				
122	Labourers transport and logistics	4,63	4,69	7,1	x				
1221	Loaders, unloaders and stock fillers	3,48	3,52	6,5	x				
1222	Garbage collectors and newspaper deliverers	1,15	1,17	10,1	x				
13	**Not elsewhere classified22**	**4,12**	**2,07**	**75,2**					**x**
131	Not elsewhere classified	4,12	2,07	75,2					x
1311	Not elsewhere classified	4,12	2,07	75,2					x
Totaal BRC-occupational classes		100	100	15,0					
Totaal BRC-occupational segments		100	100	20,4					
Totaal BRC-occupational groups		100	100	23,3					

[22] Contains records coded with occupation unknown, occupations not relevant in the Dutch labour market (e.g. 1113 Traditional chiefs and heads of villages) and nfd-codes that cannot be classified in one of the occuaptionals groups.

Appendix 3. Quality of fully automatic coding process ISCO 2008

The distribution of the coding across the occupational classes, segments and groups of ISCO 2008 according to process variant 1 and fully automatically coded according to process variant 2, the proportion per fully automatically coded category with a different code in process variant 1 and an indication of the quality, LFS 2016 unweighted total 92 thousand records.

		Process variant 2 (fully automatic Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category [23]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24]				
					Good (0-9% in-correct)	Sufficien t (10-19% in-correct)	Mediocr e (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
ISCO 2008 major group, sub-major group, minor group		%	%	%					
0	**Armed forces occupations**	**0,26**	**0,32**	**17,4**		**x**			
00	Armed forces occupations (not further defined)	0,01	0,10	50,0				x	
000	Armed forces occupations (not further defined)	0,01	0,10	50,0				x	
01	Commissioned armed forces officers	0,02	0,05	82,6					x
011	Commissioned armed forces officers	0,02	0,05	82,6					x
02	Non-commissioned armed forces officers	0,06	0,07	60,4				x	
021	Non-commissioned armed forces officers	0,06	0,07	60,4				x	
03	Armed forces occupations, other ranks	0,17	0,09	75,8					x
031	Armed forces occupations, other ranks	0,17	0,09	75,8					x
1	**Managers**	**6,44**	**5,64**	**34,9**			**x**		
10	Managers (not further defined)	1,86	0,31	96,8					x
100	Managers (not further defined)	1,86	0,31	96,8					x
11	Chief executives, senior officials and legislators	0,71	1,29	35,1			x		
111	Legislators and senior officials	0,24	0,41	38,7				x	

[23] Manual coding within a specific category is recommended when the share of unequal codes exceeds 10%, underlying categories with sufficient quality need not be coded manually.

[24] The proportion incorrect is calculated on the basis of a regression analysis into the relationship between the percentages of unequal codes and incorrect codes, as explained in appendix 1, model C.

.

		Process variant 2 (fully automatic Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category [23]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24]				
					Good (0-9% incorrect)	Sufficient (10-19% incorrect)	Mediocre (20-29% incorrect)	Bad (30-49% incorrect)	Very bad (>49% incorrect)
ISCO 2008 major group, sub-major group, minor group		%	%	%					
112	Managing directors and chief executives	0,47	0,88	36,7				x	
12	Administrative and commercial managers	2,22	1,48	45,7				x	
120	Administrative and commercial managers (not further defined)	0,01	0,91	100,0					x
121	Business services and administration managers	1,14	0,57	33,6			x		
122	Sales, marketing and development managers	1,07	1,88	59,1				x	
13	Production and specialised services managers	1,39	0,00	35,2			x		
132	Manufacturing, mining, construction, and distribution managers	0,45	0,75	40,4				x	
133	Information and communications technology service managers	0,30	0,23	46,0				x	
134	Professional services managers	0,64	0,89	30,4			x		
14	Hospitality, retail and other services managers	0,26	0,68	35,9			x		
141	Hotel and restaurant managers	0,12	0,19	27,2			x		
142	Retail and wholesale trade managers	0,06	0,37	15,5		x			
143	Other services managers	0,07	0,12	69,2					x
2	**Professionals**	**24,43**	**24,85**	**12,2**		**x**			
20	Professionals (not further defined)	0,11	0,36	12,4		x			
200	Professionals (not further defined)	0,11	0,36	12,4		x			
21	Science and engineering professionals	2,69	2,42	26,3			x		
210	Science and engineering professionals (not further defined)	0,00	0,00	100,0					x
211	Physical and earth science professionals	0,08	0,07	40,8				x	
212	Mathematicians, actuaries and statisticians	0,08	0,06	31,4			x		
213	Life science professionals	0,25	0,20	39,5				x	
214	Engineering professionals (excluding electrotechnology)	1,11	1,09	26,7			x		
215	Electrotechnology engineers	0,13	0,13	29,1			x		
216	Architects, planners, surveyors and designers	1,04	0,87	27,8			x		
22	Health professionals	3,55	3,33	17,0		x			
221	Medical doctors	1,00	0,93	14,7		x			
222	Nursing and midwifery professionals	1,54	1,38	25,2			x		
223	Traditional and complementary medicine professionals	0,04	0,05	47,1				x	
224	Paramedical practitioners	0,01	0,01	27,3			x		
225	Veterinarians	0,06	0,06	7,4	x				
226	Other health professionals	0,90	0,91	9,9	x				

ISCO 2008 major group, sub-major group, minor group		Process variant 2 (fully automatic Proportion to total %	Process variant 1 Proportion to total %	Proportion of unequal codes per fully automatically coded category [23] %	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24] Good (0-9% in-correct)	Sufficien t (10-19% in-correct)	Mediocr e (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
23	Teaching professionals	4,87	4,94	8,1	x				
230	Teaching professionals (not further defined)	0,31	0,23	76,8					x
231	University and higher education teachers	0,69	0,60	33,7			x		
232	Vocational education teachers	0,34	0,40	38,5				x	
233	Secondary education teachers	1,11	1,26	18,0		x			
234	Primary school and early childhood teachers	1,36	1,45	8,2	x				
235	Other teaching professionals	1,05	1,00	30,8			x		
24	Business and administration professionals	6,54	6,92	21,2		x			
240	Business and administration professionals (not further defined)	0,01	0,00	100,0					x
241	Finance professionals	2,09	1,84	26,7			x		
242	Administration professionals	3,01	3,43	21,7		x			
243	Sales, marketing and public relations professionals	1,42	1,65	26,4			x		
25	Information and communications technology professionals	3,07	3,16	14,4		x			
250	Information and communications technology professionals (not further defined)	0,10	0,09	67,7					x
251	Software and applications developers and analysts	2,18	2,29	16,0		x			
252	Database and network professionals	0,79	0,77	12,4		x			
26	Legal, social and cultural professionals	3,62	3,72	16,5		x			
261	Legal professionals	0,67	0,76	7,2	x				
262	Librarians, archivists and curators	0,10	0,09	29,2			x		
263	Social and religious professionals	1,56	1,69	15,7		x			
264	Authors, journalists and linguists	0,63	0,63	14,8		x			
265	Creative and performing artists	0,66	0,56	30,2			x		
3	**Technicians and associate professionals**	**14,62**	**15,37**	**19,7**		**x**			
31	Science and engineering associate professionals	2,03	2,24	27,5			x		
311	Physical and engineering science technicians	0,95	0,99	29,3			x		
312	Mining, manufacturing and construction supervisors	0,51	0,65	28,4			x		
313	Process control technicians	0,25	0,24	33,2			x		
314	Life science technicians and related associate professionals	0,06	0,09	15,8		x			
315	Ship and aircraft controllers and technicians	0,26	0,26	25,2			x		
32	Health associate professionals	2,69	2,83	15,6		x			
321	Medical and pharmaceutical technicians	0,57	0,58	16,3		x			

ISCO 2008 major group, sub-major group, minor group		Process variant 2 (fully automatic Proportion to total %	Process variant 1 Proportion to total %	Proportion of unequal codes per fully automatically coded category [23] %	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24] Good (0-9% in-correct)	Sufficien t (10-19% in-correct)	Mediocr e (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
322	Nursing and midwifery associate professionals	0,78	0,91	18,6		x			
323	Traditional and complementary medicine associate professionals		0,01						
324	Veterinary technicians and assistants	0,03	0,03	0,0	x				
325	Other health associate professionals	1,31	1,30	15,2		x			
33	Business and administration associate professionals	5,54	6,16	21,6		x			
331	Financial and mathematical associate professionals	1,11	1,22	12,7		x			
332	Sales and purchasing agents and brokers	1,31	1,64	34,3			x		
333	Business services agents	1,01	1,11	26,6			x		
334	Administrative and specialised secretaries	1,30	1,41	16,7		x			
335	Regulatory government associate professionals	0,81	0,78	28,0			x		
34	Legal, social, cultural and related associate professionals	3,74	3,51	22,2		x			
340	Legal, social, cultural and related associate professionals (not further defined)	0,00	0,00	100,0					x
341	Legal, social and religious associate professionals	2,82	2,55	23,9			x		
342	Sports and fitness workers	0,51	0,56	19,4		x			
343	Artistic, cultural and culinary associate professionals	0,41	0,39	20,5		x			
35	Information and communications technicians	0,60	0,63	35,2			x		
351	Information and communications technology operations and user support technicians	0,44	0,44	34,9			x		
352	Telecommunications and broadcasting technicians	0,17	0,14	39,1				x	
4	**Clerical support workers**	**9,62**	**9,65**	**13,2**		**x**			
40	Clerical support workers (not further defined)	0,20	0,16	44,3				x	
400	Clerical support workers (not further defined)	0,20	0,16	44,3				x	
41	General and keyboard clerks	2,05	2,52	18,6		x			
410	General and keyboard clerks (not further defined)	0,48	0,36	57,4				x	
411	General office clerks	0,84	1,53	24,8			x		
412	Secretaries (general)	0,64	0,55	17,7		x			
413	Keyboard operators	0,08	0,08	49,3				x	
42	Customer services clerks	2,34	2,08	22,1		x			
421	Tellers, money collectors and related clerks	0,18	0,15	28,0			x		
422	Client information workers	2,16	1,93	21,9		x			

		Process variant 2 (fully automatic Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category [23]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24]				
					Good (0-9% in-correct)	Sufficien t (10-19% in-correct)	Mediocr e (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
ISCO 2008 major group, sub-major group, minor group		%	%	%					
43	Numerical and material recording clerks	3,39	3,54	20,5		x			
431	Numerical clerks	1,69	1,60	22,2		x			
432	Material-recording and transport clerks	1,70	1,94	19,4		x			
44	Other clerical support workers	1,64	1,35	27,2			x		
441	Other clerical support workers	1,64	1,35	27,2			x		
5	**Service and sales workers**	**18,83**	**19,62**	**10,2**		**x**			
50	Service and sales workers (not further defined)	0,00	0,00	100,0					x
500	Service and sales workers (not further defined)	0,00	0,00	100,0					x
51	Personal service workers	5,39	5,67	8,9	x				
510	Personal service workers (not further defined)	0,01	0,00	100,0					x
511	Travel attendants, conductors and guides	0,25	0,25	12,1		x			
512	Cooks	0,56	0,66	4,6	x				
513	Waiters and bartenders	2,91	3,11	5,7	x				
514	Hairdressers, beauticians and related workers	0,72	0,78	3,0	x				
515	Building and housekeeping supervisors	0,48	0,50	32,7			x		
516	Other personal services workers	0,46	0,39	27,2			x		
52	Sales workers	7,76	8,57	10,2	x				
520	Sales workers (not further defined)	0,00	0,00	100,0					x
521	Street and market salespersons	0,04	0,08	27,5			x		
522	Shop salespersons	4,90	5,84	11,4		x			
523	Cashiers and ticket clerks	1,38	1,28	9,6	x				
524	Other sales workers	1,44	1,37	21,5		x			
53	Personal care workers	4,62	4,36	15,4		x			
530	Personal care workers (not further defined)	0,00	0,00	100,0					x
531	Child care workers and teachers' aides	1,79	1,58	21,8		x			
532	Personal care workers in health services	2,83	2,78	11,9		x			
54	Protective services workers	1,05	1,02	15,6		x			
541	Protective services workers	1,05	1,02	15,6		x			
6	**Skilled agricultural, forestry and fishery workers**	**1,47**	**1,74**	**9,4**	**x**				
61	Market-oriented skilled agricultural workers	1,45	1,72	9,1	x				
611	Market gardeners and crop growers	0,91	1,12	11,1		x			

		Process variant 2 (fully automatic Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category [23]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24]				
					Good (0-9% in-correct)	Sufficient (10-19% in-correct)	Mediocre (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
ISCO 2008 major group, sub-major group, minor group		%	%	%					
612	Animal producers	0,44	0,53	14,3		x			
613	Mixed crop and animal producers	0,10	0,05	73,4					x
62	Market-oriented skilled forestry, fishery and hunting worker	0,02	0,02	33,3			x		
621	Forestry and related workers	0,00	0,01	0,0	x				
622	Fishery workers, hunters and trappers	0,02	0,01	35,3			x		
7	**Craft and related trades workers**	**6,81**	**7,32**	**12,1**		**x**			
70	Craft and related trades workers (not further defined)	0,22	0,60	59,2				x	
700	Craft and related trades workers (not further defined)	0,22	0,60	59,2				x	
71	Building and related trades workers, excluding electricians	2,53	2,61	12,6		x			
711	Building frame and related trades workers	1,40	1,45	14,4		x			
712	Building finishers and related trades workers	0,66	0,73	15,4		x			
713	Painters, building structure cleaners and related trades workers	0,46	0,43	12,2		x			
72	Metal, machinery and related trades workers	1,91	1,92	15,6		x			
720	Metal, machinery and related trades workers (not further defined)	0,00	0,00	100,0					x
721	Sheet and structural metal workers, moulders and welders, and related workers	0,53	0,55	12,1		x			
722	Blacksmiths, toolmakers and related trades workers	0,31	0,29	25,8			x		
723	Machinery mechanics and repairers	1,07	1,09	17,7		x			
73	Handicraft and printing workers	0,32	0,29	28,5			x		
731	Handicraft workers	0,14	0,11	34,1			x		
732	Printing trades workers	0,18	0,18	24,3			x		
74	Electrical and electronic trades workers	0,79	0,87	24,1			x		
741	Electrical equipment installers and repairers	0,70	0,76	24,9			x		
742	Electronics and telecommunications installers and repairers	0,09	0,11	32,5			x		
75	Food processing, wood working, garment and other craft and r	1,04	1,04	24,6			x		
751	Food processing and related trades workers	0,52	0,49	24,9			x		
752	Wood treaters, cabinet-makers and related trades workers	0,20	0,21	11,8		x			
753	Garment and related trades workers	0,20	0,19	24,5			x		
754	Other craft and related workers	0,12	0,15	51,4				x	
8	**Plant and machine operators, and assemblers**	**3,65**	**3,93**	**12,5**		**x**			
81	Stationary plant and machine operators	0,74	0,79	39,5				x	
810	Stationary plant and machine operators (not further defined)	0,33	0,19	75,1					x

		Process variant 2 (fully automatic Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category [23]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24]				
					Good (0-9% in-correct)	Sufficient (10-19% in-correct)	Mediocre (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
ISCO 2008 major group, sub-major group, minor group		%	%	%					
811	Mining and mineral processing plant operators	0,04	0,03	58,3				x	
812	Metal processing and finishing plant operators	0,03	0,05	41,9				x	
813	Chemical and photographic products plant and machine operators	0,04	0,07	38,9				x	
814	Rubber, plastic and paper products machine operators	0,05	0,12	55,3				x	
815	Textile, fur and leather products machine operators	0,10	0,10	52,2				x	
816	Food and related products machine operators	0,05	0,13	65,2					x
817	Wood processing and papermaking plant operators	0,02	0,04	57,1				x	
818	Other stationary plant and machine operators	0,07	0,05	65,2					x
82	Assemblers	0,21	0,29	33,3			x		
821	Assemblers	0,21	0,29	33,3			x		
83	Drivers and mobile plant operators	2,70	2,85	4,4	x				
830	Drivers and mobile plant operators (not further defined)	0,03	0,39	96,8					x
831	Locomotive engine drivers and related workers	0,04	0,06	10,0	x				
832	Car, van and motorcycle drivers	0,70	0,70	12,6		x			
833	Heavy truck and bus drivers	1,35	1,03	28,0			x		
834	Mobile plant operators	0,55	0,64	5,3	x				
835	Ships' deck crews and related workers	0,03	0,03	20,0		x			
9	**Elementary occupations**	**9,99**	**10,09**	**8,3**	**x**				
91	Cleaners and helpers	3,08	3,07	8,9	x				
910	Cleaners and helpers (not further defined)	0,02	0,00	100,0					x
911	Domestic, hotel and office cleaners and helpers	2,81	2,84	8,6	x				
912	Vehicle, window, laundry and other hand cleaning workers	0,25	0,23	21,9		x			
92	Agricultural, forestry and fishery labourers	0,40	0,37	24,8			x		
921	Agricultural, forestry and fishery labourers	0,40	0,37	24,8			x		
93	Labourers in mining, construction, manufacturing and transpo	4,26	4,33	8,7	x				
931	Mining and construction labourers	0,11	0,13	22,4		x			
932	Manufacturing labourers	0,67	0,68	21,2		x			
933	Transport and storage labourers	3,48	3,52	6,5	x				
94	Food preparation assistants	1,11	1,15	7,0	x				
941	Food preparation assistants	1,11	1,15	7,0	x				
95	Street and related sales and service workers	0,01	0,01	61,5				x	

		Process variant 2 (fully automatic Proportion to total	Process variant 1 Proportion to total	Proportion of unequal codes per fully automatically coded category [23]	Quality indication for fully automatic coding on the basis of estimated proportion incorrect [24]				
					Good (0-9% in-correct)	Sufficient (10-19% in-correct)	Mediocre (20-29% in-correct)	Bad (30-49% in-correct)	Very bad (>49% in-correct)
ISCO 2008 major group, sub-major group, minor group		%	%	%					
951	Street and related service workers	0,00	0,00	0,0	x				
952	Street vendors (excluding food)	0,01	0,01	66,7					x
96	Refuse workers and other elementary workers	1,13	1,16	9,5	x				
961	Refuse workers	0,08	0,12	22,4		x			
962	Other elementary workers	1,05	1,05	8,6	x				
99	Unknown	3,88	1,47	81,4					x
99	Unknown	3,88	1,47	81,4					x
999	Unknown	3,88	1,47	81,4					x
Totaal ISCO-major groups		100	100	16,8					
Totaal ISCO-submajor groups		100	100	21,5					
Totaal ISCO-minor groups		100	100	24,4					

Appendix 4. Crosstabs 1st aggregation level BRC 2014 and ISCO 2008

Table 1. Distribution per fully automatically coded occupational class of BRC 2014 across occupational classes coded according to process variant 1, LFS 2016, more than 92 thousand records unweighted

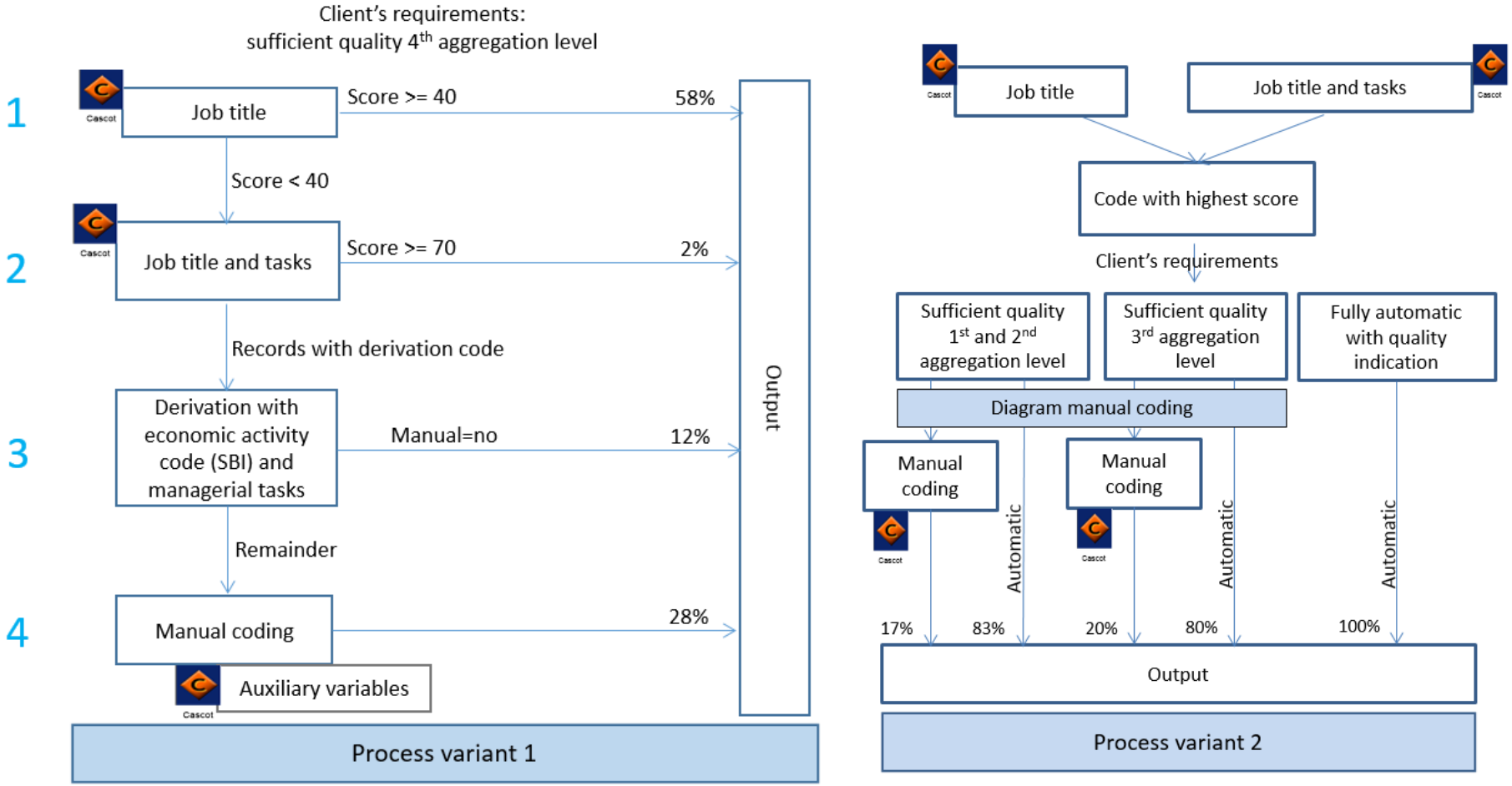
	Process variant 1													
	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
Fully automatic, process variant 2	%													
1- Pedagogical occupations	90,4	0,2	0,3	1,7	0,7	0,4	0,8	0,2	0,1	3,4	0,9	0,1	0,8	100
2- Creative and linguistic occupations	1,6	77,3	3,5	4,1	2,0	0,4	5,2	2,5	0,2	0,9	0,5	0,2	1,7	100
3- Commercial occupations	0,1	0,3	89,4	3,1	1,6	0,5	1,1	0,4	0,3	0,4	1,3	1,0	0,6	100
4- Business economics and administrative occupations	0,4	0,3	2,9	87,9	1,4	1,1	2,2	0,8	0,2	1,0	0,6	0,5	0,9	100
5-Managers	0,6	0,5	14,1	8,2	64,0	1,4	3,5	2,0	1,1	0,8	1,9	0,6	1,4	100
6- Public administration, security and legal occupations	0,5	0,2	1,3	5,6	1,3	83,9	2,2	0,7	0,1	1,9	0,6	0,7	1,1	100
7- Technical occupations	0,4	0,4	1,7	2,5	0,8	0,5	86,2	0,8	0,7	0,9	1,3	1,2	2,8	100
8- ICT occupations	0,3	0,8	1,3	4,2	1,3	0,6	4,0	85,2	0,0	0,7	0,3	0,2	1,1	100
9- Agricultural occupations	0,2	0,1	1,7	1,4	0,3	0,2	2,2	0,0	91,0	1,0	0,8	0,8	0,4	100
10- Care and welfare occupations	1,9	0,1	0,4	2,1	0,7	0,7	1,5	0,3	0,1	90,3	0,8	0,2	1,0	100
11- Service occupations	0,4	0,1	1,1	1,0	0,5	0,4	1,0	0,1	0,5	1,0	92,9	0,4	0,6	100
12- Transport and logistics occupations	0,1	0,1	1,2	0,9	0,1	0,5	1,3	0,1	0,1	0,2	0,7	94,4	0,4	100
13- Not elsewhere classified	1,9	1,4	9,7	13,9	9,7	2,4	17,2	2,9	2,8	4,4	5,4	3,5	24,8	100
Total	7,1	2,1	11,9	18,7	5,2	3,3	12,8	3,8	2,1	13,1	10,0	7,8	2,1	100

Table 2. Distribution per fully automatically coded ISCO 2008 major group across major groups codied according to process variant 1, LFS 2016, more than 92 thousand records unweighted

	Process variant 1											
	0	1	2	3	4	5	6	7	8	9	99	total
Fully automatic, process variant 2	%											
0-Armed forces occupations	82,6	0,9	2,1	3,4	0,4	6,8	0,0	1,3	0,9	0,0	1,7	100
1-Managers	0,1	65,1	11,9	10,7	1,3	6,2	1,1	1,5	0,6	0,3	1,3	100
2-Professionals	0,1	1,9	87,8	4,9	1,2	1,4	0,2	0,9	0,3	0,3	0,9	100
3-Technicians and associate professionals	0,3	1,4	7,7	80,3	2,6	3,6	0,2	1,4	0,8	0,4	1,2	100
4-Clerical support workers	0,1	0,6	2,3	4,5	86,8	3,1	0,1	0,7	0,4	0,8	0,7	100
5-Service and sales workers	0,0	1,2	2,0	3,2	1,0	89,8	0,2	0,4	0,2	1,4	0,4	100
6-Skilled agricultural, forestry and fishery workers	0,1	0,4	0,7	1,4	1,1	2,0	90,6	1,0	0,7	1,7	0,3	100
7-Craft and related trades workers	0,1	0,4	1,6	2,1	0,6	2,4	0,4	87,9	2,2	1,6	0,6	100
8-Plant and machine operators, and assemblers	0,1	0,1	0,7	1,7	0,9	1,9	0,4	3,5	87,5	2,8	0,6	100
9-Elementary occupations	0,0	0,1	0,5	0,8	0,8	3,3	0,7	0,8	0,9	91,7	0,4	100
99-Onbekend	0,1	11,3	17,9	10,1	4,6	13,4	2,4	11,3	4,8	5,3	18,6	100
Total	0,3	5,6	24,9	15,4	9,7	19,6	1,7	7,3	3,9	10,1	1,5	100

Appendix 5. Schematic overview and practical examples

Schematic representation of the two process variants, percentages based on table 2 and table 7.



Practical examples

The examples below aim to illustrate a variety of ways in which answers on occupation and tasks may pass through the two processes, it is not the intention to present an exhaustive overview.

Example 1

Job title: interieurverzorger (housekeeper, cleaning lady)

Tasks: cleaning of the interior

Process sequence variant 1

The record passes through step 1 and is coded occupation unknown with score 0. Code occupation unknown is the result of Cascot's conclusion-rules. After combining with tasks, the record, with score 57, matches with index entry ’interieur verzorger huishouden’ ISCO code 9111 ‘Domestic cleaners and helpers’. This score, however, is lower than 70, which is the threshold value used in step 2. No derivation code has been assigned in steps 1 and 2, with the effect that the record does not pass through step 3 and is then coded manually in step 4. The coder decides in this case on the basis of the additional auxiliary variables whether or not nfd-code 9110 is applicable, or whether the person cleans in either private households (9111) or in hotels, offices (9112).

Process sequence variant 2

The record passes through step 1, and with score 69, it matches with index entry ’interieur verzorger huishouden’, ISCO code 9111 ‘Domestic cleaners and helpers’. The record then passes through step 2 after the combination with the tasks and with score 56 a match is made with index-entry ‘interieur verzorger huishouden', ISCO code 9111. The score of the coding in step 1, being higher than in step 2, is consequently copied to the output with the corresponding score. As the score is higher than 40, there is no further need for manual coding, even if the client has indicated to opt for manual coding to raise the quality. In this case, although it is unknown whether the person cleans in private households, or in a company or an institution, the coding is correct at the 3rd aggregation level of ISCO (minor group 911 ‘Domestic, hotel and office cleaners and helpers’) or BRC 2014 (occupational group 1121 ‘Cleaners’).

Example 2

Job title: Voorbereidend medewerker (preparatory assistant)

Tasks: preparing vegetables and washing dishes

Process sequence variant 1

The record passes through step 1, and with score 33, it matches with index entry 'box medewerker' ISCO code 4321 ‘Stock clerks'. This score is lower than 40, therefore the record goes on to step 2. Adding the tasks in step 2 results in a match with index entry 'afwasmedewerker’ (dishwashing assistant) with score 61 and ISCO code 9412 'Kitchen helpers'. A diversion code was not assigned in the first two coding steps, and the score in the second step is less than 70. Consequently, the record proceeds to step 4 to be coded manually with the same code as assigned in step 2.

Process sequence variant 2

The process is similar to variant 1, but in this variant the code that was recorded in step 2 was taken over in the output, since the score is higher than recorded in step 1 and also higher than 40. There is no further need for manual coding.

Example 3

Job title: Accountmanager

Tasks: Obtaining insight in demand from new and existing customers

Process sequence variant 1

The record goes through step 1 and is assigned a derivation code with score 99. This is the result of Cascot's default code rules. In the second step the tasks are combined with the job title and with score 26 there is a match with index entry accountmanager pensioenen ' ISCO code 3321 ' Insurance representatives'. This is below 70, and because a diversion code was assigned in the first step, the derivation rules are followed in step 3. Given the fact that this person is employed in SBI code 6202 ‘Computer consultancy activities’, ISCO code 2434 ‘Information and communications technology sales professionals’ is assigned in this step. This code ends up in the output.

Process sequence variant 2

The record goes through step 1 and with score 38 a match is found with index entry 'account manager retail' ISCO code 3322 'Representatives, account managers retail and export managers'. This score and the index entry found result from the Cascot rules. In the second step, after combination with the tasks with score 26, a match is found with index entry ‘accountmanager pensioenen ', ISCO code 3321 ‘Insurance representatives’. This score is lower than the score of the first step, so the code recorded in the first step ends up in the output. If the choice was made to use manual coding to enhance the quality, this record will be forwarded to manual coding, because as shown in appendix 3, within minor group 332 the

share of unequal codes in process variant 1 exceeds 10%, and also in occupational group 0321 'Representatives and purchasers' of BRC 2014 to which this code is mapped, the proportion is higher than 10%. During manual coding, ISCO code 2434 is assigned based on the information on the economic activity (SBI).

Note. This approach is selected for several occupations that derive to an ISCO code in process variant 1 through the derivation rules based on available auxiliary variables. In process variant 2, these occupations are assigned the most probable code based on the job description, but the score is also (if necessary) assigned a value lower than 40, so that these records can be identified for manual coding in case higher quality is required.

Example 4

Job title: Opticien

Tasks: Perform eye measurements and providing advice

Process sequence variant 1

Based on the description of the occupation in step 1 with score 97 the record matches with index entry 'opticien', ISCO code 3254 ‘Optician’. The score is higher than the threshold value and the record proceeds to the output.

Process sequence variant 2

Based on the job description in step 1 with score 97 the record matches with index entry 'opticien' the ISCO code 3254 ‘Dispensing opticians’. In the second step, the activities are combined with the job title and score 52 matches the same index entry 'opticien' and ISCO code 3254 ‘Dispensing opticians’. The coding with the highest score of the first step is taken over in the output.