Deliverable Number: D8.18
Deliverable Title: Workshop to present SERISS socio-economic coding tools
Work Package: 8 'A coding module for socio-economic survey questions'

Deliverable type: Other
Dissemination status: Public

Submitted by: ESS ERIC (HQ/CITY)
Author: Elena Sommer (ESS HQ/CITY)
Date Submitted: October 2017

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654221.
SERISS (Synergies for Europe’s Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe’s social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe’s grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Sommer, E. (2017) Workshop to present SERISS socio-economic coding tools. Deliverable 8.18 of the SERISS project funded under the European Union’s Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables
Explanatory note

The workshop ‘Harmonised coding for socio-economic survey questions’ took place on 4th-5th September in Amsterdam. The main purpose of the workshop was to introduce coding tools developed under SERISS to a wider group of researchers, survey practitioners (e.g. cross-national survey infrastructures, commercial survey agencies, representatives of non-profit organisations conducting social surveys) and other stakeholders (e.g. national statistics institutes, employment agencies) involved in designing, coding and analysing socio-economic questions, to offer them an opportunity to try out the tools during the workshop and to provide feedback and suggestions for tool upgrades and training materials. In total 41 participants were present at the workshop representing research, practice and policy making fields.

The workshop was a joint event with the ‘SERISS Survey Experts Network’ organised under SERISS Work Package 5. This enabled us to hold a two-day workshop, providing enough time for tool demonstrations and practical sessions, and facilitated attendance by a larger, more diverse group of participants. To facilitate an exchange between participants, the workshop had an interactive format with presentations followed by practice sessions and time for questions as well as a general discussion session at the end of the workshop. To encourage post-workshop communication between the participants, a digital group was set up on Basecamp where participants can share information and feedback.

This deliverable provides documentation of the workshop consisting of the workshop agenda, participant list and copies of presentations given during the workshop. A further report summarising the presentations and the discussion outcomes is available as SERISS Deliverable 5.10 ‘Survey Network Meeting report 2’.

Accompanying materials

- Workshop agenda
- Participant list
- Presentations
  - ‘Introduction to SERISS WP 8 - A coding module for socio-economic survey questions’, by Kea Tijdens, University of Amsterdam
  - ‘Challenges of matching’, by Rita Bergenhill, Swedish Public Employment Service
  - ‘The SERISS occupation, industry and employment status database’, by Kea Tijdens, University of Amsterdam
  - ‘An occupation>>industry prediction for the job coder’, by Michelle Belloni, Ca’ Foscari University of Venice
  - ‘Name generator in household surveys’, by Ella Schwartz, Hebrew University of Jerusalem
  - ‘Computer Assisted Measurement and Coding of Educational Qualifications in Multicultural Surveys - Presentation of the CAMCES/SERISS education tool’, by Verena Ortmanns and Silke Schneider, GESIS - Leibniz Institute for the Social Sciences
  - ‘Surveycodings.org’, by Maurice Martens, CentERdata
  - ‘Development of CASCOT 5.5 (a multi-language text coding tool)’, by Peter Elias, Warwick Institute for Employment Research
  - ‘CASCOT in practice – example of Statistics Netherlands’, by Sue Westerman, Dutch Central Bureau of Statistics (CBS)
## Workshop agenda

**SERISS Survey Experts Workshop ‘Harmonised coding for socio-economic survey questions’**  
De Burcht, Henri Polaklaan 9, 1018 CP Amsterdam  
4th-5th September 2017

**Day 1, 4th September 2017, Chair: Eric Harrison (ESS ERIC, CITY, University of London)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:30 - 13:30</td>
<td>Registration and lunch</td>
</tr>
<tr>
<td>13:30 – 14:40</td>
<td><strong>Introduction to the workshop</strong></td>
</tr>
<tr>
<td></td>
<td>• 13:30 – 13:45 Introduction to the SERISS project &amp; workshop content</td>
</tr>
<tr>
<td></td>
<td>Eric Harrison, European Social Survey (ESS ERIC), City, University of London</td>
</tr>
<tr>
<td></td>
<td>• 13:45 – 14:00 Relevance of high-quality measure of socio-economic variables</td>
</tr>
<tr>
<td></td>
<td>Peter Elias, University of Warwick</td>
</tr>
<tr>
<td></td>
<td>• 14:00 – 14:15 Introduction to SERISS Work Package “A coding module for socio-economic survey questions”</td>
</tr>
<tr>
<td></td>
<td>Kea Tijdens, University Amsterdam</td>
</tr>
<tr>
<td></td>
<td>• 14:15 – 14:40 Self-introduction of participants</td>
</tr>
<tr>
<td>14:40 – 15:40</td>
<td><strong>Coding occupation, industry and employment status</strong></td>
</tr>
<tr>
<td></td>
<td>• 14:40 – 15:00 Challenges of matching</td>
</tr>
<tr>
<td></td>
<td>Rita Bergenhill, Swedish Public Employment Service</td>
</tr>
<tr>
<td></td>
<td>• 15:00 – 15:20 Presentation of the SERISS occupation, industry and employment status database</td>
</tr>
<tr>
<td></td>
<td>Kea Tijdens, University Amsterdam</td>
</tr>
<tr>
<td></td>
<td>• 15:20 – 15:40 An occupation &gt;&gt; industry prediction for the job coder</td>
</tr>
<tr>
<td></td>
<td>Michelle Belloni, Ca’ Foscari University of Venice</td>
</tr>
<tr>
<td>15:40 – 16:00</td>
<td><strong>COFFEE BREAK</strong></td>
</tr>
<tr>
<td>16:00 – 16:20</td>
<td><strong>Measuring social networks in surveys</strong></td>
</tr>
<tr>
<td></td>
<td>• 16:00 – 16:20 Name generator in household surveys</td>
</tr>
<tr>
<td></td>
<td>Ella Schwartz, Hebrew University of Jerusalem</td>
</tr>
<tr>
<td>16:20 – 17:30</td>
<td><strong>Practical session</strong></td>
</tr>
<tr>
<td></td>
<td>• 16:20 – 16:40 Presentation of the SERISS module demo</td>
</tr>
<tr>
<td></td>
<td>Maurice Martens, CentERdata</td>
</tr>
<tr>
<td></td>
<td>• 16:40 – 17:10 Module in practice – participants try out SERISS occupation, industry and employment status module &amp; name generator tool</td>
</tr>
<tr>
<td></td>
<td>• 17:10 – 17:30 Questions and answers</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>09:00 – 10:20</td>
<td>Coding educational attainment and field of education</td>
</tr>
<tr>
<td></td>
<td>• 09:00 – 09:30 Presentation of the SERISS education database</td>
</tr>
<tr>
<td></td>
<td>Verena Ortmanns, GESIS - Leibniz Institute for the Social Sciences</td>
</tr>
<tr>
<td></td>
<td>• 09:30 – 10:00 Module in practice – training session</td>
</tr>
<tr>
<td></td>
<td>• 10:00 – 10:20 Questions and answers</td>
</tr>
<tr>
<td>10:20 – 10:40</td>
<td>COFFEE BREAK</td>
</tr>
<tr>
<td>10:40 – 11:30</td>
<td>Module benefits for survey holders</td>
</tr>
<tr>
<td></td>
<td>• 10:40 – 11:00 Use of module in different survey modes</td>
</tr>
<tr>
<td></td>
<td>Maurice Martens, CentERdata</td>
</tr>
<tr>
<td></td>
<td>• 11:00 – 11:20 Module in practice – training session</td>
</tr>
<tr>
<td></td>
<td>• 11:20 – 11:30 Questions and answers</td>
</tr>
<tr>
<td>11:30 – 12:20</td>
<td>Batch coding using CASCOT: Computer Assisted Structured Coding Tool</td>
</tr>
<tr>
<td></td>
<td>• 11:30 – 11:50 Introduction to multilingual CASCOT software</td>
</tr>
<tr>
<td></td>
<td>Peter Elias, University of Warwick</td>
</tr>
<tr>
<td></td>
<td>• 11:50 – 12:10 CASCOT in practice – example of Statistics Netherlands</td>
</tr>
<tr>
<td></td>
<td>Sue Westerman, Dutch Central Bureau of Statistics (CBS)</td>
</tr>
<tr>
<td></td>
<td>• 12:10 – 12:20 Questions and answers</td>
</tr>
<tr>
<td>12:20 – 13:20</td>
<td>LUNCH</td>
</tr>
<tr>
<td>13:20 – 13:50</td>
<td>Variable Harmonisation Software</td>
</tr>
<tr>
<td></td>
<td>• 13:20 – 13:40 Introduction to variable harmonisation software CharmStats</td>
</tr>
<tr>
<td></td>
<td>Kristi Winters, GESIS - Leibniz Institute for the Social Sciences</td>
</tr>
<tr>
<td></td>
<td>• 13:40 – 13:50 Questions and answers</td>
</tr>
<tr>
<td>13:50 – 14:50</td>
<td>Discussion session: Harmonisation challenges &amp; future of coding</td>
</tr>
<tr>
<td></td>
<td>Chair: Peter Elias, University of Warwick</td>
</tr>
<tr>
<td>14:50 – 15:00</td>
<td>Workshop closure &amp; AOB</td>
</tr>
</tbody>
</table>
List of workshop participants

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution/Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rui Alves</td>
<td>Statistics Portugal (INE)</td>
</tr>
<tr>
<td>Eric Balster</td>
<td>CentERdata, Tilburg University</td>
</tr>
<tr>
<td>Michele Belloni</td>
<td>Ca’ Foscari University of Venice</td>
</tr>
<tr>
<td>Rita Bergenhill</td>
<td>Swedish Public Employment Service</td>
</tr>
<tr>
<td>Yvette Boodhna</td>
<td>Ipsos MORI</td>
</tr>
<tr>
<td>Claudia Brunori</td>
<td>European Values Study (EVS), Tilburg University</td>
</tr>
<tr>
<td>Bernhard Christoph</td>
<td>Institute for Employment Research (IAB)</td>
</tr>
<tr>
<td>Julius De Zeeuw</td>
<td>Netherlands Employment Office (UWV)</td>
</tr>
<tr>
<td>Peter Elias</td>
<td>University of Warwick</td>
</tr>
<tr>
<td>Tom Emery</td>
<td>Generations and Gender Programme (GGP), Netherlands</td>
</tr>
<tr>
<td>Rebeka Falle Zorman</td>
<td>Interdisciplinary Demographic Institute (NIDI)</td>
</tr>
<tr>
<td>Martin Friedichs</td>
<td>GESIS – Leibniz Institute for Social Sciences</td>
</tr>
<tr>
<td>Harry Ganzeboom</td>
<td>International Social Survey Programme (ISSP), University Amsterdam</td>
</tr>
<tr>
<td>Eric Harrison</td>
<td>European Social Survey (ESS ERIC), City, University of London</td>
</tr>
<tr>
<td>Daniel Horn</td>
<td>Development Pathways</td>
</tr>
<tr>
<td>Agata Kaczmarek-Firth</td>
<td>European Union Statistics on Income and Living Conditions (EU-SILC), Eurostat</td>
</tr>
<tr>
<td>Simon King</td>
<td>RTI International</td>
</tr>
<tr>
<td>Sophie König</td>
<td>Statistics Austria</td>
</tr>
<tr>
<td>Piergiorgio Lovaglio</td>
<td>Interuniversity Research Centre for Public Services (CRISP), University of Milano-Bicocca</td>
</tr>
<tr>
<td>Ruud Luijkx</td>
<td>European Values Study (EVS), Tilburg University</td>
</tr>
<tr>
<td>Angelica Maineri</td>
<td>European Values Study (EVS), Tilburg University</td>
</tr>
<tr>
<td>Maurice Martens</td>
<td>CentERdata, Tilburg University</td>
</tr>
<tr>
<td>Simona Mineo</td>
<td>Institute for the Development of Vocational Training of Workers (ISFOL)</td>
</tr>
<tr>
<td>Verena Ortmanns</td>
<td>GESIS – Leibniz Institute for Social Sciences</td>
</tr>
<tr>
<td>Paulien Osse</td>
<td>WageIndicator, University Amsterdam</td>
</tr>
<tr>
<td>Agnes Parent-Thirion</td>
<td>European Working Conditions Survey, Eurofound</td>
</tr>
<tr>
<td>Anja Perry</td>
<td>GESIS – Leibniz Institute for Social Sciences</td>
</tr>
<tr>
<td>Anita Pugliese</td>
<td>Gallup</td>
</tr>
<tr>
<td>Zbigniew Sawiński</td>
<td>Harmonization Project, Polish Academy of Sciences</td>
</tr>
<tr>
<td>Ella Schwartz</td>
<td>Israel Gerontological Data Center, Hebrew University of Jerusalem</td>
</tr>
<tr>
<td>Elena Sommer</td>
<td>European Social Survey (ESS ERIC), City, University of London</td>
</tr>
<tr>
<td>Stephanie Stuck</td>
<td>Survey of Health, Ageing and Retirement in Europe (SHARE ERIC), Munich Center for the Economics of Aging (MEA)</td>
</tr>
<tr>
<td>Kea Tijdens</td>
<td>WageIndicator, University Amsterdam</td>
</tr>
<tr>
<td>Ursula Till-Tentschert</td>
<td>European Union Agency for Fundamental Rights</td>
</tr>
<tr>
<td>William Thorn</td>
<td>Programme for the International Assessment of Adult Competencies (PIAAC), OECD</td>
</tr>
<tr>
<td>Karl Viilmann</td>
<td>Statistics Estonia</td>
</tr>
<tr>
<td>Irena Vipavc Brvar</td>
<td>Slovenian National Data Archive, University of Ljubljana</td>
</tr>
<tr>
<td>Sue Westerman</td>
<td>Dutch Central Bureau of Statistics (CBS)</td>
</tr>
<tr>
<td>Aleksandra Wilczynska</td>
<td>Kantar Public</td>
</tr>
<tr>
<td>Kristi Winters</td>
<td>GESIS – Leibniz Institute for Social Sciences</td>
</tr>
<tr>
<td>Charlie Wroth-Smith</td>
<td>UK Office for National Statistics (ONS)</td>
</tr>
</tbody>
</table>
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654221.

Introduction to WP 8 - A coding module for socio-economic survey questions

Kea Tijdens, University of Amsterdam & WageIndicator web-survey

SERISS workshop
4th September 2017

WP 8 Overall objectives

• To ensure cross-country harmonised and high-quality coding of core socio-economic variables
  – Occupation >> ISCO-08
  – Industry >> NACE Rev. 2
  – Education and field of education >> ISCED-11
  – Employment and socio-economic status >> ISCE-93/ ESEG-14
• To prepare coding tools for the measurement of these variables, departing from technologies used for web-surveys, and aiming for respondents’ self-identification
• To build an online survey module for the tools
• To adapt the tools for CAPI surveys
• To build a name generator for social network measurement
Why an online survey module?

• Core variables mostly asked using open-ended questions
• Requiring office coding
  -->> time-consuming & expensive
  -->> coding problems: answers can be coded at same level of detail & some are unidentifiable
• Lack of cross-country harmonization
• -->> The coding module has to solve these problems

WP 8 - 8 tasks

| Task 8.1 | Programming the survey module |
| Task 8.2 | Compile API-database of occupations |
| Task 8.3 | Compile API-databases of educational attainment + field |
| Task 8.4 | Compile the API-database of industries |
| Task 8.5 | Compile the API-database of employment status |
| Task 8.6 | Design related survey questions&answers |
| Task 8.7 | Consultation and dissemination |
| Task 8.8 | Design a name generator |
WP 8 – 6 partners

• UVA --> occupation, industry and employment status plus related survey tool
• ESS --> training, workshops, dissemination
• GESIS --> education tool
• SHARE/MUNICH --> name generator
• SHARE/CENTERDATA --> software programming
• SHARE/UNIVE--> occupation>industry prediction

Presentations in this workshop

• TASK 8.1
  • The SERISS module demo
• TASK 8.2, 8.4, 8.5
  • Databases of occupations, industries and employment status
• TASK 8.2
  • An occupation >> industry prediction for the ‘job coder’
• TASK 8.3
  • Presentation of the SERISS/CAMCES education database
• TASK 8.8
  • Name generator in household surveys
• Thank you for your attention

• See https://seriss.eu/about-seriss/work-packages/TASK8-a-coding-module-for-socio-economic-survey-questions/
Challenges of matching

SERISS Workshop
Harmonised coding of socio-economic survey questions
Amsterdam

2017-09-04
Rita Bergenhill
rita.bergenhill@arbetsformedlingen.se

Arbetsförmedlingen

Challenges of matching

Matching - what does the word mean?
Challenges of matching

Employer

Demands
- Skills/know how
- Working experience
- Education

Offers
- Occupational role
- Place of work
- Salary
- Full-time/part-time

Job seeker

Demands
- Occupational role
- Place of work
- Salary
- Full-time/part-time

Offers
- Skills/know how
- Working experience
- Education

Confusion on terms used in matching

job
occupation
Challenges of matching

Definition of job and occupation (ILO)

A job is defined as a set of tasks and duties performed or meant to be performed by a person who is employed or self-employed.

The concept of occupation refers to the type of work performed. A number of different jobs with similar tasks and duties constitutes an occupation.

Example: A job can consist of driving a bus between airport terminals and aircraft, another of driving a coach between the airport and the city center, while a third might be driving a school bus.

Three separate jobs with slightly different conditions but the occupation is the same - bus driver.

Challenges of matching

Three realities at the labour market
Challenges of matching

The reality of employers
skills need expressed by using job titles or/and company-specific titles

The reality of jobseekers
Range of skills expressed by using education titles, personal titles or old occupation names that are not longer used at labour market

The reality of PES
matching vacancies and job seekers by using classification of occupations and standard list of occupation names

Challenges of matching

Job ads

• difficulties to express the real skills need

• focus on describing the company’s business rather than on the tasks in the current job and on what skills are demanded to be able to perform the tasks

job seekers do not understand the requirements
employers do not get suitable candidates
job seekers do not find relevant jobs
Challenges of matching

Method for identifying required skills/competences – by Stockholm County Council

- Identify roles/occupations and produce definitions for the roles/occupations
- Identify tasks for the roles/occupations they have identified
- Identify the "real" skills/competences required to perform the tasks (key words)
- Produce definitions for the skills/competences for the role/occupation
- Produce definitions for formal skills/competences

Challenges of matching

Register quality - impact on statistical information

- Job seeker’s profiles tend to be incomplete or even incorrect
- Job ads tend to be connected to incorrect unit groups of occupations
Challenges of matching

All the challenges >> reliability of big data analysis?

- User friendly tools for employers and job seekers
- In addition to IT support, methods are also needed

Thank you for your attention!
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654221.

The SERISS occupation, industry and employment status database

Kea Tijdens, University of Amsterdam & WageIndicator web-survey

SERISS workshop
4th September 2017

Outline

• Introduction
• The database of occupations
• The database of industries
• The database of socio-economic and employment status
Outline

• Introduction
• The database of occupations
• The database of industries
• The database of socio-economic and employment status

WP8 Overall objectives

• WP8: A coding module for socio-economic survey questions
• Long-list survey questions (occupation, industry) are mostly asked open-ended, requiring office-coding
• In web-surveys they can be asked in closed format, inviting respondents to self-select an answer from a database
• Respondents browse the database using search tree and/or autosuggest
• WP8: Tasks include building these databases
Outline

• Introduction
• The database of occupations
• The database of industries
• The database of socio-economic and employment status

Coding practices in multi-country surveys

– multi-country coding indexes do not exist
  >> black box: are the same occupations coded similarly across countries?
– national survey agencies mostly perform their own coding in ISCO-08, sometimes 4 digit, mostly 2 digit
– only few examples of surveys, where national coding was tested
– WageIndicator web-survey on work and wages in 92 countries uses a multilingual database of occupations
Validating occupational coding

• Two viewpoints
  – occupations differ so much across countries that coding practices in multi-country surveys cannot be validated
  – occupations do not differ and national coding can be validated (compare to industry NACE classification with Eurostat’s efforts in harmonising coding)

• How to validate occupational coding across countries?
  – occupational titles cannot be translated beyond ISCO-4-digit, --> countries have different occupational titles
  – occupational titles can be translated in a globalising world, due to
    • pressure towards cross-country standardization, e.g. QESH auditor
    • global equipment suppliers -> tasks in jobs become similar
    • globalisation of the economy -> need to understand occupations
    • pressure on labour market mobility in Europe > need to understand occ.

Explore both viewpoints in WP8

• Expand WageIndicator database of occupations
  – from 1,700 to 4,000 occupations, all coded ISCO-08
  – translate these titles into 47 languages for 99 countries

• Merge coding indexes from nat. statistical offices
  – use translation software to test to what extent the occupations in the indexes are coded similarly
  – --> add these occupations to the database
Merging coding indexes

• Coding indexes
  – collected from 34 countries (mostly from NSOs), 20 could be used
    • because they were from non-English speaking countries
    • because index was coded ISCO-08
    • because index was beyond ISCO-08 4 digit
    • because index was not incomplete
  – merging occupational titles resulted in 70,000 records
• Are these occupations coded similar across countries?
  – all titles from the 20 indexes were translated into English, using Google translate & online dictionaries
  – if same translation & same ISCO code, the titles were assumed to be the same occupational title (remains still 55,000 records)
  – currently further identifying same occupations & reducing records
  – result: validity analysis + additions to the database

Tasks in SERISS WP8

• Make occupation database available for survey-holders
  – program an API for web-surveys on desktop, tablet, smartphone, plus an interface for CAPI surveys
• In web-surveys: Respondents self-identify occupation
  – through a search tree (IPod menu) or
  – through an autosuggest box (Google search type)
• In CAPI: Interviewers identify respondent’s occupation
  – through an autosuggest box (Google search type)
  – turn the screen and ask respondents to self-identify
Outline

• Introduction
• The database of occupations
• The database of industries
• The database of socio-economic and employment status

The Industry survey question

• “In which industry do you work?”
  – Typically asked in as open-ended survey question
  – Question difficult to answer without interviewer
  – Question is cognitive demanding
    • In web-surveys with open-ended questions respondents tend to answer the name of their company
    • In web-surveys with closed questions respondents tend to skip it
• A multi-lingual industry database
  – WageIndicator web-survey asks this question in closed format, using an industry database with 321 entries
  – Coded NACE Rev 2: -->> 4-digit codes, except manufacturing industry -->> 3-digit codes
Tasks in SERISS WP8

• Expand the WageIndicator industry database
  – to 99 countries and 47 languages
  – make it available to survey holders, using an API

• Develop an occupation>industry prediction
  – improve data quality and reduce respondent time
  – --> develop an occupation>industry prediction

Outline

• Introduction
• The database of occupations
• The database of industries
• The databases of socio-economic and employment status
Database of employment status

- International Classification by Status in Employment (ICSE-93)
  - maintained by ILO, and the 2013 draft revisions
- Categories in ICSE93
  - 100 Employee
  - 200 Employer
  - 300 Self-employed
  - 400 Member of producers' cooperatives
  - 500 Contributing family workers
  - 600 Other forms of employment not elsewhere classified
- Extra details in draft 2013
  - Employees on permanent versus temporary employment, agency work
  - Self-employed, including measurement of bogus employment
  - Measurement of informal work
- Survey questions and answers
  - In WP8 survey questions and answers designed to measure this classification in web-surveys
  - The questions and answers are available for the 99 countries and the 47 languages

Database of socio-economic status

- European Socio-Economic Groups (ESeG-2014)
  - Early 2000s Eurostat -> European Socio-Economic Classification (ESeC)
  - Revision needed because of ISCO-08
- ESeG-2014
  - two-level classification of 9 groups and 42 subgroups, four variables:
    - Occupation (ISCO08)
    - employment status (employee / self-employed)
    - people not in paid employment (retired / student / disabled)
    - age
- SERISS WP8.5
  - For measuring ESeG-2014 WP8 designed survey questions and answers, translated for 99 countries
• Thank you for your attention

• See https://seriss.eu/about-seriss/work-packages/wp8-a-coding-module-for-socio-economic-survey-questions/

• See http://www.wageindicator.org/main
An occupation >> industry prediction for the job coder

Michele Belloni, A. Brugiavini, E. Buia

Ca’ Foscari University of Venice

and

K. Tijdens, RA: A. Zerain

AIAS-UvA

SERISS Survey Experts Workshop ‘Harmonised coding for socio-economic survey questions’

Amsterdam
4-5 September 2017

• deliverable D8.11 of WP8 and its Task 8.4: “Compile the API-database of industries” (European Commission, Directorate-General For Research & Innovation, Research infrastructure, 2015, ANNEX 1)

• Responsible partners: University Ca’ Foscari of Venice and University of Amsterdam (UvA)

• General aim of the work: facilitate the coding of industries during the interview

• Alternative approaches: a database of company names; occupational titles database providing respondents an alert once they had selected an unlikely occupation-industry combination (high number of unlikely combinations and risk of drop-out)
What we did

- We report the most likely industries (NACE 2.0) given the 4-digit ISCO-08 occupation as coded by the jobcoder
- The “occupation>industry prediction” allows to reduce substantially the lookup list of industries and highlights the most relevant ones for the respondent
- This shortlist always include the option “other”: in that case the full industry lookup table is shown

The LPM model

We estimate a set of $k$ linear probability models (LPM) – one for each ISCO 08 4-digits group ($k=1\ldots436$)

$$Y_{it}^k = \sum_{i,j} \beta_j D_{\text{NACE}i,j} + \varepsilon_i$$

- $Y_{i}=1$ if individual $i$ reported ISCO code $k$; 0 otherwise
- $D_{\text{NACE}}=1$ if individual $i$ reported NACE industry $j$; 0 otherwise
- $\varepsilon$ is a i.i.d. error term
- $\beta_j$ is the estimated marginal effect. It can be directly interpreted as the probability that an individual reporting a given occupation $k$ works in industry $j$

- In addition to the $k=1\ldots436$ models, we estimated many additional models for major (10 out of 10), sub-major (34 out of 43) and minor (92 out of 130) ISCO 08 groups - depending on data availability. These additional models are not used by the Jobcoder
Dataset: requirements

- ISCO 08 up to 4-digits (436 groups) jointly with Industry 2-digits NACE version 2.2 (88 groups)
- **Country coverage**: need to include as many countries as possible among those in wp. This required merging multiple sources of data.
- Need to use only recent data: change in occupation/industry classification structure (e.g.: in Wave 5 of ESS occupations were coded in ISCO 88). Conversion tables do not guarantee one-to-one correspondence and we preferred not to introduce assumptions.
- **Main need**: huge number of observations because we need to estimate a model for each ISCO unit groups (i.e. 436 models)

Dataset: composition

We end up with merging two types of datasets:

- **CAPI surveys**: ESS Wave 6-7 and EWCS Wave 5 (2010; «restricted version»: the public release only shows 2-digits)
- **Web surveys**: mostly the Wageindicator dataset
- Final dataset after cleaning: 1,114,240 observations
  - About 10% from CAPI -> using Web surveys was crucial
  - Almost all the 99 countries covered (38 by the CAPI surveys; only 12 are not covered)
## Retrieved surveys and their respective sample size: estimation sample

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CAPI SURVEYS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESS 6</td>
<td>46,012</td>
<td>4.13</td>
<td>4.13</td>
</tr>
<tr>
<td>ESS 7</td>
<td>25,536</td>
<td>2.29</td>
<td>6.42</td>
</tr>
<tr>
<td>EWCS 2010 RES</td>
<td>43,589</td>
<td>3.91</td>
<td>10.33</td>
</tr>
<tr>
<td><strong>WEB SURVEYS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monsterempl</td>
<td>4,571</td>
<td>0.41</td>
<td>10.74</td>
</tr>
<tr>
<td>t-online</td>
<td>27,346</td>
<td>2.45</td>
<td>13.20</td>
</tr>
<tr>
<td>wageindicator</td>
<td>458,926</td>
<td>41.19</td>
<td>54.39</td>
</tr>
<tr>
<td>wageindicator-empl</td>
<td>72,136</td>
<td>6.47</td>
<td>60.86</td>
</tr>
<tr>
<td>wageindicator-journalists</td>
<td>12</td>
<td>0.00</td>
<td>60.86</td>
</tr>
<tr>
<td>wageindicator-self</td>
<td>3,872</td>
<td>0.35</td>
<td>61.21</td>
</tr>
<tr>
<td>wageindicator-unem</td>
<td>514</td>
<td>0.05</td>
<td>61.25</td>
</tr>
<tr>
<td>wageindicator-wicare</td>
<td>66</td>
<td>0.01</td>
<td>61.26</td>
</tr>
<tr>
<td>websurvey2015</td>
<td>3,193</td>
<td>0.29</td>
<td>61.55</td>
</tr>
<tr>
<td>wiltie</td>
<td>230,931</td>
<td>20.73</td>
<td>82.27</td>
</tr>
<tr>
<td>wiltie-empl</td>
<td>188,337</td>
<td>16.90</td>
<td>99.17</td>
</tr>
<tr>
<td>wiltie-self</td>
<td>4,447</td>
<td>0.40</td>
<td>99.57</td>
</tr>
<tr>
<td>wiltie-unem</td>
<td>3,39</td>
<td>0.30</td>
<td>99.88</td>
</tr>
<tr>
<td>wiltie201301-empl</td>
<td>7</td>
<td>0.00</td>
<td>99.88</td>
</tr>
<tr>
<td>wiltie201301-self</td>
<td>2</td>
<td>0.00</td>
<td>99.88</td>
</tr>
<tr>
<td>wipaper</td>
<td>1,353</td>
<td>0.12</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1,114,240</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

## Results: top 20 industries for ISCO 2310

<table>
<thead>
<tr>
<th>Industry, NACE Rev.2 2-digit</th>
<th>coefaster</th>
<th>University and higher education teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>72, Scientific research and development</td>
<td>0.322***</td>
<td>0.127***</td>
</tr>
<tr>
<td>85, Education</td>
<td>0.0244***</td>
<td>0.0157***</td>
</tr>
<tr>
<td>99, Activities of extraterritorial organisations and bodies</td>
<td>0.0203***</td>
<td>0.0153***</td>
</tr>
<tr>
<td>21, Manufacture of basic pharmaceutical products and pharmaceuti</td>
<td>0.0166***</td>
<td>0.0149***</td>
</tr>
<tr>
<td>75, Veterinary activities</td>
<td>0.0157***</td>
<td>0.0128***</td>
</tr>
<tr>
<td>3, Fishing and aquaculture</td>
<td>0.0153***</td>
<td>0.0122***</td>
</tr>
<tr>
<td>91, Libraries, archives, museums and other cultural activities</td>
<td>0.0149***</td>
<td>0.0122***</td>
</tr>
<tr>
<td>2, Forestry and logging</td>
<td>0.0128***</td>
<td>0.0118***</td>
</tr>
<tr>
<td>70, Activities of head offices; management consultancy activitie</td>
<td>0.0118***</td>
<td>0.0112***</td>
</tr>
<tr>
<td>90, Creative, arts and entertainment activities</td>
<td>0.0112***</td>
<td>0.01***</td>
</tr>
<tr>
<td>73, Advertising and market research</td>
<td>0.01***</td>
<td>0.00838***</td>
</tr>
<tr>
<td>84, Public administration and defence; compulsory social securit</td>
<td>0.0072***</td>
<td>0.00577***</td>
</tr>
<tr>
<td>74, Other professional, scientific and technical activities</td>
<td>0.00576**</td>
<td>0.00537***</td>
</tr>
<tr>
<td>86, Human health activities</td>
<td>0.00576**</td>
<td>0.00518***</td>
</tr>
<tr>
<td>96, Other personal service activities</td>
<td>0.00518***</td>
<td>0.00518***</td>
</tr>
</tbody>
</table>
Results: full set

- See [excel file](#)

Sensitivity: Country heterogeneity

- A country specific estimation for each ISCO 08 unit group would not be possible due to lack of sufficient data for many non-EU countries.
- We made a simple clustering exercise: we split the 38 countries included in the CAPI surveys into two groups: the EU countries and the other countries which, supposedly have a different labor market and possibly are characterised by different associations between occupation and industry.
- We leave out the web surveys because its inclusion would have mixed up group of country differences with the different type of sampling (random vs volunteer).
- We compared the results picking up some ISCO unit groups at random: The estimated most likely industries were very similar, in some case in a different order, among the two groups of countries.
Other issues

The possibility to account for additional covariate such as gender, age or education was also initially considered.

- AIAS prepared a huge merged dataset with harmonized information also on other variables such as:
  1. Public Sector: dummy question on whether respondent works in public sector
  2. Supervisory position: dummy question on whether respondent has a supervisory position
  3. Number of supervisees
  4. Firm size: organised in three categories
  5. Paid work: dummy question on whether respondent has a paid job
  6. Self-employed: dummy question on whether respondent is self-employed
  7. Gender
  8. Year of birth
  9. .....

Other issues

This dataset can be very useful for various other studies in our case it could have imposed limits:

- These info must be collected earlier in the CAPI: too restricted (note that we already require that occupation is asked before industry)
- Not enough variability is left in some cases by including many covariates
- Number of observations is lower due to missing values
Personal Social Networks

- Personal social networks are important in people’s lives
  - Related to aspects such as health, cognition, depression and mortality
  - Modify the effects of stressors on health and wellbeing
  - At different points in life and under differing circumstances shape the paths to key outcomes
- Not easy to measure
- Various ways to measure social networks in surveys, e.g.
  - Indirect approach
  - Direct approach (name generator)
Name Generator Approach

- Obtain information about specific persons within older adults' personal networks
- Subjective mapping of social networks and their characteristics (e.g. size, composition)
- Used in few aging survey:
  - LASA (Longitudinal Aging Study Amsterdam)
  - NSHAP (National Social Life, Health, and Aging Project)
  - SHARE

Survey of Health, Ageing and Retirement in Europe

- Multidisciplinary, cross-national panel study
- Social, economic and health situation
- Target group 50+ population
- Face to face CAPI
- First wave in 2004
- 7th wave is now in the field with 28 countries
Name Generator Approach

- Used in SHARE in 2011 (W4) and 2015 (W6)

- Ask respondents to **name** persons that are important for them
  - Up to seven network members
  - Follow up questions on SN members & relationship characteristics
  - Overall satisfaction with network (or not having a network)

- Network persons are linked with other questionnaire modules - children module, social support and financial transfers (=>very complex routing)

Baseline SN Module

- Follow-up Questions
  - Gender
  - Geographical Proximity
  - Frequency of Contact
  - Emotional Closeness
  - Overall Satisfaction with Network
Going Longitudinal

- Baseline assessment of actual social network
  - Including follow up questions

- Compare both networks
  - Using preloaded information from previous social network
  - Link SN members between waves
  - Ask for reason if SN member “got lost”

The Longitudinal Dimension

Wave 4 (preloaded)

| Maria (sister) | Pascal (son) | Clara (daughter) |

Reason why Maria is not mentioned again?

Clara mentioned again?

Wave 6

| Pascal (son) | Clara mentioned again? |
|----------------|
| "lost" member |

Additional Follow-up Questions added in w6:
- Year of birth
- Employment Status
- Relationship Status

Follow-up Questions (both waves):
- Gender
- Geographical Proximity
- Frequency of Contact
- Emotional Closeness
- Overall Satisfaction with Network

New member:

Thomas (friend)
Linkage of SN members

- CAPI instrument displays both lists of networks members
- Interviewer goes through the list of preloaded ‘old’ SN members asking if they have been mentioned again this time
- Name of w6 network member disappears from the screen if chosen

“Lost” SN Members

- If a network person was not mentioned again we ask:
  What is the main reason you didn't mention Maria this time?
  1. I forgot, Maria should have been included
  2. I moved
  3. Maria moved
  4. Maria died
  5. I became ill or had a health problem
  6. Maria became ill or had a health problem
  7. Respondent does not recognize the named person
  8. We are no longer close
  9. Wrong, Maria WAS mentioned this time
  97. Other reason (specify)
CAPI example

What do we get?

- We see overall network changes like:
  - Size of social network
  - Network composition
  - Quality of relations
  - Support given to or provided by network members
  - etc
- We also see individual changes within the network composition and reasons for changes
### Summary of name interpreter questions

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spousenet2</td>
<td>Is there a spouse in the social network?</td>
</tr>
<tr>
<td>famnet / childnet / siblingnet / parentnet / friendnet / formalnet / othernet</td>
<td>Count of different role relationship categories in a social network</td>
</tr>
<tr>
<td>womennet / mennet</td>
<td>Count of women/men in a social network</td>
</tr>
<tr>
<td><em>prx</em></td>
<td>Information on geographical distance to network members</td>
</tr>
<tr>
<td><em>contact</em></td>
<td>Information on contact frequencies with network members</td>
</tr>
<tr>
<td><em>close</em></td>
<td>Information on emotional closeness to network members</td>
</tr>
<tr>
<td>year_mean</td>
<td>Average year of birth of network members</td>
</tr>
<tr>
<td><em>fin</em> / <em>gift</em> / <em>care</em></td>
<td>Information on given or received financial / personal help to / from network members</td>
</tr>
<tr>
<td>sn_satisfaction</td>
<td>Satisfaction with personal network (1-10)</td>
</tr>
<tr>
<td>sn_scale</td>
<td>Scale of social connectedness, a summary scale of the social network data (higher = higher connectedness)</td>
</tr>
</tbody>
</table>

### What do we get?

#### SN characteristics in Waves 4 and 6

<table>
<thead>
<tr>
<th></th>
<th>Wave 4</th>
<th>Wave 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social network size</td>
<td>2.5</td>
<td>2.7</td>
</tr>
<tr>
<td>Proportion kin</td>
<td>77.6</td>
<td>79.8</td>
</tr>
<tr>
<td>Mean contact frequency</td>
<td>6.1</td>
<td>6.1</td>
</tr>
<tr>
<td>Mean emotional closeness</td>
<td>3.2</td>
<td>3.3</td>
</tr>
<tr>
<td>Satisfaction with the social network</td>
<td>8.9</td>
<td>8.8</td>
</tr>
</tbody>
</table>

#### Proportion reporting social network changes in Waves 4 and 6

- Decrease
- No change
- Increase
Panel Variables

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>w4_sn_mentioned_again_x</td>
<td>Was w4 SN member X mentioned again in w6?</td>
</tr>
<tr>
<td>w6_sn_mentioned_before_x</td>
<td>Was w6 SN member X mentioned before in w4?</td>
</tr>
<tr>
<td>w6_sn_w4_position_x</td>
<td>What was the position of w6 SN member X in the w4 SN?</td>
</tr>
<tr>
<td>panel_lost</td>
<td>Count of lost wave 4 SN members</td>
</tr>
<tr>
<td>panel_new</td>
<td>Count of new wave 6 SN members</td>
</tr>
<tr>
<td>panel_continued</td>
<td>Count of w4 SN members that were mentioned again in w6</td>
</tr>
</tbody>
</table>

Most panel respondents experienced some internal turnover

- 2+ SN members: 63%
- 1 SN member: 37%
- No such SN member: 30%
- 20% respondents have no such SN member

<table>
<thead>
<tr>
<th>Stable SN members</th>
<th>Lost SN members</th>
<th>New SN members</th>
</tr>
</thead>
<tbody>
<tr>
<td>38%</td>
<td>60%</td>
<td>35%</td>
</tr>
<tr>
<td>43%</td>
<td>30%</td>
<td>28%</td>
</tr>
<tr>
<td>19%</td>
<td>40%</td>
<td>37%</td>
</tr>
</tbody>
</table>
Thank You!

Questions?
Outline

• Project background and aims

• Components of the CAMCES tool

• Empirical evaluation: Results from the IAB-SOEP Migration Sample

• Future developments and implementations
Project background and aims

Project background I

- Education is...
  - the most central background variable in surveys
  - very difficult to measure, especially cross-culturally

- Nationally:
  - Educational systems become increasingly differentiated
    - Short lists incomplete, difficult for respondents
    - Long lists difficult to administer
  - Migration results in foreign qualifications
    - Often inappropriate instruments used, not taking context of origin into account
Project background II

• Cross-nationally:
  – Idiosyncratic institutional differences across countries, translation of names of qualifications impossible
  – Thus use of country-specific items and output-harmonisation
  – Education data not consistently coded across surveys
    • Ortmanns and Schneider, 2016a and 2016b
  – ISCED levels lack cross-national validity
    • Schneider 2010, Müller and Klein 2008, Kerckhoff et al. 2002...

➢ Education classifications difficult to implement
➢ Substantial logistics of ex-ante output harmonisation
➢ No standards for surveys of migrants
State of research

- Many cross-national survey data harmonize education variables using International Standard Classification of Education (ISCED)

**ISCED 1997 main levels**

- ISCED 0: Pre-primary education
- ISCED 1: Primary education
- ISCED 2: Lower secondary
- ISCED 3: Upper secondary education
- ISCED 4: Post-secondary, non-tertiary education
- ISCED 5: First stage of tertiary education
- ISCED 6: Second stage of tertiary education
State of research

- Many cross-national survey data harmonize education variables using International Standard Classification of Education (ISCED)
- Conclusion: Thus data should be consistent

However, prior research revealed discrepancies of harmonised education distributions across surveys:
  - France: EU-LFS and ESS (Kieffer, 2010)
  - European countries: EU-LFS, EU-SILC, ESS (Schneider, 2009)
  - European countries: ESS, EVS, ISSP, Eurobarometer (Ortmanns and Schneider, 2016a) and EU-LFS, EU-SILC, PIAAC (Ortmanns and Schneider, 2016b)
Comparing education variable across surveys

Duncan’s Dissimilarity Index:

- Compares education distributions between EU-LFS and the other surveys
- Interpretation: Percentage of cases that would have to change categories in order to achieve equal distributions

![Education distribution in the Netherlands](image)

D = 42%

www.seriss.eu
@SERISS_EU
How can we explain these discrepancies?

Main reasons for inconsistencies:

- Errors due to the measurement instrument
  - Showcards contains ambiguous and vague terms
  - Incomplete showcard
- Errors in data processing
  - Accidental missclassifications into ISCED
  - Intended deviations from official ISCED mappings
CAMCES project aim: develop new tools

- consisting of survey measurement instruments (brief question module), database and software interface
- that enable the accurate, detailed and cross-nationally comparable measurement, coding and harmonisation of highest educational qualification obtained
- in computer-assisted surveys (where content can be shown visually to respondents, i.e. CAPI, CAWI)

CAMCES project

- Project funding: German Leibniz Association
- Project duration: 2013 to 2017
- Project team:
  - Silke L. Schneider
  - Jessica Herzing / Roberto Briceno-Rosas
  - Verena Ortmanns
- Programming of CAMCES tool:
  - CAPI tool: Gerald Mahlmeister
  - Web tool, website, database backend: CentERdata, esp. Maurice Martens
Components of the CAMCES tool

- CAMCES question module
- CAMCES survey interfaces:
  - Combination Box (CB), including algorithm for textstring matching
  - Search Tree (ST)
- CAMCES/SERISS database
CAMCES question module

- Three main elements:
  - Introduction (concept clarification, scope)
  - Identify relevant educational system (“education context”)
  - Ask about highest educational qualification obtained

- Optional elements
  - Country-specific school-leaving certificates
  - Loop to capture more than one qualification
  - Year of qualification, years of education...

- Translations (TRAPD) from DE into EN, NL, ES, FR(, RU)

Introduction (example: UK)

“Now some questions about your formal education. This covers schooling, higher education and officially recognized vocational education. Formal education excludes on-the-job training, training programmes for the unemployed (e.g. through a Jobcentre or the Department for Work and Pensions), as well as continuing training lasting less than 6 months full-time.”
Default Context – all respondents

- Where did you obtain your highest educational qualification?
  - Survey country - context 1
  - Survey country - context 2
  - ...
  - Abroad

- Example response categories for surveys in the UK:
  - United Kingdom - England, Wales and Northern Ireland
  - United Kingdom - Scotland
  - Abroad

Detailed context for respondents educated abroad

- Where did you obtain your highest educational qualification?

- Morocco
- Nepal
- Netherlands
- New Zealand
- Nicaragua
- Nigeria
- Norway
- Oman
- Pakistan
- Panama
- Paraguay
- Peru
- Philippines
- Poland
CAMCES survey interfaces

1. Combination Box (CB) - in CAPI only when latin alphabet
   - text entry box for search in DB
   - algorithm to show most relevant results
   - selection of best match by respondent
   - may not always be successful (memory, motivation)
   - requires more instructions than ST

2. Search Tree (ST)
   - equivalent of show card but more detailed
   - nesting of qualifications in broader categories
   - used as fallback for respondents unsuccessful in CB
   - may entail more social desirability or ‘rounding’
   - may be slower (requires more reading)
Combination Box (CB) - Italy

What is your highest educational qualification?

- Please enter the qualification in Italian.
- Please do not type the subject area or your occupation.
- Please select the best match.

Search Tree (ST) - Italy

What is your highest educational qualification?
This project has received funding from the research and innovation programme under grant agreement No 654221.

Search Tree (ST) - Russia

What is your highest educational qualification?

- Вообще не учился в школе, незаконченная начальная школа
- Аттестат об основном общем образовании
  - Аттестат об основном общем образовании, 7 классов [по системе до 1958 года]
  - Аттестат об основном общем образовании, 8 классов [по советской системе 60-80-х годов]
  - Аттестат об основном общем образовании, 9 классов [по современной системе]
- Начальное или среднее профессионально-техническое образование или на базе основного общего образования
- Аттестат о законченном среднем общем образовании
- Начальное или среднее профессионально-техническое образование или на базе полного среднего
- Высшее образование

Search Tree (ST) - Iraq

What is your highest educational qualification?

- لم يذهب إلى المدرسة
- أقل من شهادة التعليم الابتدائي
- شهادة التعليم الابتدائي
- شهادة التعليم المتوسط
- شهادة التعليم الإعدادي
- شهادة التعليم المهني
- معهد علي
- جامعة
- درجة بكالوريوس (5 سنة)
- درجة بكالوريوس (4 سنة)
- بكالوريوس الطب (3 سنة)
- دبلوم عالي
- درجة ماجستير
- درجة دكتوراه
Loop filter question

- Do you have yet another educational qualification?
  - This can be as high or lower.
  - Please remember not to mention on-the-job training, training programmes for the unemployed (e.g. through a Jobcentre or the Department for Work and Pensions), nor continuing training lasting less than 6 months full-time.

  □ Yes  → same context yes/no? If no: back to detailed context, if yes: back to qualification

  □ No  → optional questions/exit

CAMCES/SERISS database

- Contexts (countries, regions/languages) of education
- Historical and current qualifications
- Classification codes and references to classifications for harmonization
  – Allows deriving syntax for recoding detailed qualifications
- Search tree structure
- “Synonyms”
  – More general and more specific terms to improve Combination Box matching
  – Linked to one or more qualifications
- Translations of terms used in the tool (e.g. “Search”)
Empirical evaluation: Results from the IAB-SOEP Migration Sample

Overview of CAMCES Pretest and Pilot Studies

2. Focus groups with SOEP interviewers (2014)
3. CAP/SI Pretest for IAB-SOEP Migration Sample (2014)
4. Quantitative pretests and pilot studies (2014-2016)
   - SOEP Innovation Sample, DE only (2014)
   - IAB-SOEP Migration Sample, only foreign (2015 and 2016)
   - CentERdata’s LISS Panel, NL only (2016)
5. Qualitative studies: cognitive interviews
   - Cross-cultural, with eye-tracking, DE (2015)
   - Think aloud, Venezuela (2016)
Overview of CAMCES Pretest and Pilot Studies

2. Focus groups with SOEP interviewers (2014)
3. CAP/SI Pretest for IAB-SOEP Migration Sample (2014)
4. Quantitative pretests and pilot studies (2014-2016)
   - SOEP Innovation Sample, DE only (2014)
   - IAB-SOEP Migration Sample, only foreign (2015 and 2016)
   - CentERdata’s LISS Panel, NL only (2016)
5. Qualitative studies: cognitive interviews
   - Cross-cultural, with eye-tracking, DE (2015)
   - Think aloud, Venezuela (2016)

Migrant Survey: IAB-SOEP Migration Sample

- Cooperation between Institute of Employment Research (IAB, Nuremberg) and German Socio-Economic Panel (SOEP, Berlin)
- Individuals who migrated to Germany since 1994, or whose parents had done so
- First wave including ~2700 households in 2013
- Oversampling by country of origin: Poland, Romania, former Soviet Union countries, Turkey, former Yugoslavia, Italy, Spain, and Greece
- Input-harmonized education measures inspired by German educational system
IAB-SOEP-MS education measures I

- Where did you last attend school?
- If abroad:
  - For how many years did you attend school?
  - With what kind of certificate did you complete your schooling?
    - Left school without school-leaving certificate
    - Completed compulsory school with certificate
    - Completed secondary school („Weiterführende Schule“) with certificate

IAB-SOEP-MS education measures II

- Have you completed vocational training or higher education in a country other than Germany? Yes/No
- If yes: What kind of education or training was this?
  - I was trained ‚on the job‘
  - I have completed a longer training/apprenticeship in a company
  - I attended a vocational school
  - I attended higher education with rather practical focus
  - I attended higher education with rather theoretical focus
  - I have completed doctoral studies
  - Other
Main Descriptives

- IAB-SOEP Migration Sample (SOEP-MS, only foreign, 2nd interview), 2015: CAP/SI
  - CB with ST as fallback (for qualifications in latin script), N=414
  - ST only (for non-latin script), N=399

- Quality indicators to look at
  - Response formats and item-nonresponse/,other'
  - Reasons for failed database lookup
  - Highest qualification reported in first loop?
  - Validation of response formats with panel information (established closed questions)?

% Item nonresponse (no answer, blank)

Established: School
Established: Post-school
DB lookup (latin)
ST fall-back
ST only (non-latin)

0 % 5 % 10 % 15 % 20 % 25 %
% other/not automatically coded

- Established: School
- Established: Post-school
- DB lookup (latin)
- ST fall-back
- ST only (non-latin)

Reasons for non-matching CB entries

- Language does not conform to context
- Occupation or field of education/training
- Qualification in DB not selected
- Educational institution or generic level
- Qualification/synonym not in DB
- Non-formal education or dropout
This project has received funding from the research and innovation programme under grant agreement No 654221.

Usage of loops in %

- **Latin script - CB+ST**
  - 2nd loop valid: 20%
  - 3rd loop valid: 80%

- **Non-latin script - ST only**
  - 2nd loop valid: 10%
  - 3rd loop valid: 90%

% Reporting highest qualifications per loop

- **Latin script - CB+ST**
  - 1st loop: 80%
  - 2nd loop: 20%
  - 3rd loop: -

- **Non-latin script - ST only**
  - 1st loop: -
  - 2nd loop: -
  - 3rd loop: 100%
% consistently coded cases compared to input-harmonized measures

CB (latin script), matched

ST fall-back after CB

ST only (non-latin script)

% inconsistently coded cases compared to established measures

CB (latin script), matched

ST fall-back after CB

ST only (non-latin script)
Summary

- Adaptive education measurement in surveys is feasible
- Many respondents skip the CB
- Many respondents do not conceptually differentiate between education, occupation and field of education/training
- Some respondents do not first report highest qualification
- Consistency with established measures looks reasonable; IAB-SOEP instrument is not ideal for comparison analysis

Future developments and implementations
Current implementations

- CAMCES tool implementation:
  - IAB-SOEP Migration Sample 2017
  - BAMF-IAB-SOEP Survey of Refugees in Germany 2017

- CAMCES/SERISS database used:
  - ReGES project - Refugees in the German Educational System hosted at the Leibniz Institute for Educational Trajectories (LifBi)
  - Partly: EVS 2017 coordinated at Tilburg University
  - Partly: new countries in SHARE round 7

Future work in SERISS WP 8.3

- Add more countries to the database
  - CAMCES covers European countries only -> extension to 99 countries covering all parts of the world

- Fields of education and training will be added
  - Adaptation of the interfaces for fields of education and training
References I


References II


Schneider, S. L. (2010). Nominal comparability is not enough: (in-)equivalence of construct validity of cross-national measures of educational attainment in the European social survey. Research in Social Stratification and Mobility, 28, 343–357. doi: 10.1016/j.rssm.2010.03.001

Thank you!

Contact: verena.ortmanns@gesis.org
Responses to a coding question

- Closed
- Open
- Something in between

What is your occupation?

What is your occupation?
- Factor
- VOLUME
- Genre
- name
- Reed
How to integrate?

- It depends
  - Connectivity
  - Software
  - Preferences

How to integrate?

- Connectivity:
  - Do you have an internet connection available during an interview?
  - Can you call [https://api.surveycodings.org](https://api.surveycodings.org)?
Surveycodings.org

- Software:
  - Can your software make calls to another server
  - Can your software make calls to a local library
  - Can your software activate a local application
  - Can you load in a list in your software?

- [https://api.surveycodings.org/codings/search.json?type=education&mode=country](https://api.surveycodings.org/codings/search.json?type=education&mode=country)

  - Will give a list back in JSON.
  - Is a service
  - You can use various parameters to formulate your request

```json
{}
```

JSON
No connectivity

- Can’t connect to the service
- Connect to tool on the same device
- Connect to webservice on local device

Problem to formulate solutions, it all depends on the software you use.

Lists

- At [https://surveyscodings.org](https://surveyscodings.org) you will be able to download the lists in excel and ddi format, your programmer should load this in your survey software
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221.

Feedback

- Thank you
- Some problems detected
- Some ideas

- Occupation Tree is only two level?
- Occupation Tree is not the ISCO classification
- There are some spelling mistakes
- Our internal code should not show
- How to deal with Gender?
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654221.

Feedback

- No industry tree?
- My list is better, can I upload it?
- Export to DDI
- Missing classifications

THANKS!

Maurice
Coding text to a classification

- Coding is the process of categorising the range of all possible answers to a pre-defined set of categories.
- The full set of categories is termed a classification. Examples are:
  - SOC 2010 (UK Standard Occupational Classification 2010)
  - ISCO 08 (International Standard Classification of Occupations 2008)
  - SIC 2007 (UK Standard Industrial Classification 2007)
- Three parts to a classification: the structure, the index and the classification rules
In total 369 Unit Groups

Structure

In total 27,738 index entries
Manual coding procedures

- Manual methods
  - code books;
  - temporary labour;
  - query resolution systems.
- No standardised approach, major variations between institutions, companies, etc. in quality of coding.
- Time-consuming, expensive.
Development of software

CASCOT: Java text coding software for any operating system.

CASOC was *ad hoc* development, funded from sales revenue.
CASCOT initially funded by ESRC, now funded via sales income.

Occupational coding in practice

- Quality of coding reflects quality of text available for coding.
- Need rules which specify how to deal problems such as ambiguous job titles (*e.g.* engineer, teacher).
- Need to be aware that machine coding of text can introduce bias.
- Need to establish ‘trade off’ between accuracy and cost.
Coding with Cascot

- Cascot will provide:
  - A list of recommendations.
  - Code, title, best matching index entry, and certainty score
- Certainty Score
  - Approximates the probability that the recommended code is correct.
  - This is represented by a number in the range 0-100.
  - People never 100% right. Computer can’t be 100% right.

Type job title

Press enter, or click ‘Code’ button
**Recommendations Table**

<table>
<thead>
<tr>
<th>Code</th>
<th>Group title</th>
<th>Best matching index entry</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1223</td>
<td>Restaurant and catering establishment managers</td>
<td>Manager (catering)</td>
<td>46</td>
</tr>
<tr>
<td>5223</td>
<td>Metal working production and maintenance fitters</td>
<td>Engineer, catering</td>
<td>31</td>
</tr>
<tr>
<td>4159</td>
<td>Other administrative occupations n.e.c.</td>
<td>Administrator, catering</td>
<td>30</td>
</tr>
<tr>
<td>3546</td>
<td>Conference and exhibition managers and organisers</td>
<td>Organiser, catering</td>
<td>30</td>
</tr>
<tr>
<td>7129</td>
<td>Sales related occupations n.e.c.</td>
<td>Representation, catering</td>
<td>30</td>
</tr>
</tbody>
</table>

**Classification Structure**

```
543 Food Preparation and Hospitality Trades
- 5431 Butchers
- 5432 Baker and flour confectioners
- 5433 Fishmongers and poultry dressers
- 5434 Chefs
- 5435 Cooks
5436 Catering and her managers
```

**Output**

5436 Catering and her managers

File: No output file
No automatic processing
### Index Entries

**5436 Catering and bar managers**

<table>
<thead>
<tr>
<th>Code</th>
<th>Group Title</th>
<th>Best Matching Index Entry</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Caterer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manager, unit (catering)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adviser, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consultant, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contractor, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Controller, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manager, airline</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Superintendent, catering</td>
<td></td>
</tr>
</tbody>
</table>

### Output

**5436 Catering and bar managers**

<table>
<thead>
<tr>
<th>Code</th>
<th>Group Title</th>
<th>Best Matching Index Entry</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Caterer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manager, unit (catering)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adviser, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consultant, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contractor, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Controller, catering</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manager, airline</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Superintendent, catering</td>
<td></td>
</tr>
</tbody>
</table>
Best recommendation selected automatically

Select another by clicking a different line

OR: Change selection via structure

Accept the selection
Large scale coding

An experienced person can code manually about 100 occupations/hour maintaining a good level of quality

What if they need to code a file of 100,000 occupational texts?

Use input/output files and automated coding

Using files

- Instead of typing every job title in, we can read them from a file.

- Rather than having to copy the code produced by Cascot we can have Cascot save the codes (and other output) to a file.
Example: Using Files

Input file (tab delimited).

<table>
<thead>
<tr>
<th>ID</th>
<th>JobTitle</th>
</tr>
</thead>
<tbody>
<tr>
<td>101001</td>
<td>3D Animator</td>
</tr>
<tr>
<td>101002</td>
<td>Account Director</td>
</tr>
<tr>
<td>101003</td>
<td>Account Executive</td>
</tr>
<tr>
<td>101004</td>
<td>Account Manager</td>
</tr>
<tr>
<td>101005</td>
<td>Accountant</td>
</tr>
<tr>
<td>101006</td>
<td>Accounts assistant</td>
</tr>
<tr>
<td>101007</td>
<td>Accounts Assistant</td>
</tr>
<tr>
<td>101008</td>
<td>Accounts Clerk</td>
</tr>
<tr>
<td>101009</td>
<td>Accounts Executive</td>
</tr>
<tr>
<td>101010</td>
<td>Accounts Manager</td>
</tr>
<tr>
<td>101011</td>
<td>Accounts Processing clerk</td>
</tr>
<tr>
<td>101012</td>
<td>Accounts Trainee</td>
</tr>
<tr>
<td>101013</td>
<td>Acoustic Engineer</td>
</tr>
<tr>
<td>101014</td>
<td>Actuarial Consultant</td>
</tr>
<tr>
<td>101015</td>
<td>Actuarial Trainee</td>
</tr>
<tr>
<td>101016</td>
<td>admin assistant</td>
</tr>
<tr>
<td>101017</td>
<td>admin assistant</td>
</tr>
<tr>
<td>101018</td>
<td>Admin officer</td>
</tr>
<tr>
<td>101019</td>
<td>Admin worker</td>
</tr>
<tr>
<td>101020</td>
<td>Administrator</td>
</tr>
</tbody>
</table>
Choose Output Items

Click Edit
Example: Using Files

Output file (Output items = "Input Record, Code, Title, Score")

<table>
<thead>
<tr>
<th>ID</th>
<th>SubTitle</th>
<th>Code</th>
<th>Title</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Animator</td>
<td>3411</td>
<td>Artists</td>
<td>60</td>
</tr>
<tr>
<td>101</td>
<td>Account Director</td>
<td>1312</td>
<td>Marketing and sales directors</td>
<td>38</td>
</tr>
<tr>
<td>101</td>
<td>Account Executive</td>
<td>3542</td>
<td>Business sales executives</td>
<td>39</td>
</tr>
<tr>
<td>101</td>
<td>Account Manager</td>
<td>3538</td>
<td>Financial accounts managers</td>
<td>30</td>
</tr>
<tr>
<td>101</td>
<td>Accountant</td>
<td>3411</td>
<td>Chartered and certified accountants</td>
<td>39</td>
</tr>
<tr>
<td>101</td>
<td>Assistant</td>
<td>4122</td>
<td>Book-keepers, payroll managers and</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>wages clerks</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Assistant</td>
<td>4122</td>
<td>Book-keepers, payroll managers and</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>wages clerks</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Clerk</td>
<td>4122</td>
<td>Book-keepers, payroll managers and</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>wages clerks</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Executive</td>
<td>3543</td>
<td>Marketing associate professionals</td>
<td>53</td>
</tr>
<tr>
<td>101</td>
<td>Manager</td>
<td>3538</td>
<td>Financial accounts managers</td>
<td>79</td>
</tr>
<tr>
<td>101</td>
<td>Processing Clerk</td>
<td>4122</td>
<td>Book-keepers, payroll managers and</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>wages clerks</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Trainee</td>
<td>4122</td>
<td>Book-keepers, payroll managers and</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>wages clerks</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Acoustic Engineer</td>
<td>3419</td>
<td>Construction and building trades n.o.c.</td>
<td>71</td>
</tr>
<tr>
<td>101</td>
<td>Actuarial Consultant</td>
<td>3425</td>
<td>Actuaries, economists and statisticians</td>
<td>57</td>
</tr>
<tr>
<td>101</td>
<td>Actuarial Trainee</td>
<td>3425</td>
<td>Actuaries, economists and statisticians</td>
<td>60</td>
</tr>
<tr>
<td>101</td>
<td>Admin assistant</td>
<td>4159</td>
<td>Other administrative occupations n.o.c.</td>
<td>39</td>
</tr>
<tr>
<td>101</td>
<td>Admin assistant</td>
<td>4159</td>
<td>Other administrative occupations n.o.c.</td>
<td>39</td>
</tr>
<tr>
<td>101</td>
<td>Admin officer</td>
<td>4159</td>
<td>Other administrative occupations n.o.c.</td>
<td>39</td>
</tr>
<tr>
<td>101</td>
<td>Admin worker</td>
<td>4159</td>
<td>Other administrative occupations n.o.c.</td>
<td>84</td>
</tr>
<tr>
<td>101</td>
<td>Administrator</td>
<td>6132</td>
<td>Pensions and insurance clerks and assistants</td>
<td>32</td>
</tr>
</tbody>
</table>

Automated coding

- Rather than choosing manually the best recommendation every time we can automate the process

- Automation options
  - fully manual
  - semi-automatic, select the certainty level (manual coding when score is below the level)
  - fully automatic
A fully automated run

- How good is this?

- Example:
  - 19,087 unique job titles
  - Coded fully automatically, sorted descending by certainty score
  - Selections from output file shown
  - Wrong codes coloured with light orange
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text</strong></td>
<td><strong>Code</strong></td>
<td><strong>Index</strong></td>
<td><strong>Title</strong></td>
<td><strong>Score</strong></td>
</tr>
<tr>
<td>2501 credit insurance broker</td>
<td>3532</td>
<td></td>
<td>Broker, insurance</td>
<td>Brokers</td>
</tr>
<tr>
<td>2502 oil executive</td>
<td>3539</td>
<td></td>
<td>Executive, company, oil</td>
<td>Business and related associ</td>
</tr>
<tr>
<td>2503 computer sales</td>
<td>3542</td>
<td></td>
<td>Salesman, computer</td>
<td>Business sales executives</td>
</tr>
<tr>
<td>2504 double glazing salesman/a common</td>
<td>3542</td>
<td></td>
<td>Salesman, glazing, double</td>
<td>Business sales executives</td>
</tr>
<tr>
<td>2509 manual</td>
<td>3995</td>
<td></td>
<td>Instructor, manual</td>
<td>Vocational and industrial</td>
</tr>
<tr>
<td>2506 booker keeper</td>
<td>4122</td>
<td></td>
<td>Bookkeeper</td>
<td>Bookkeepers, payroll man</td>
</tr>
<tr>
<td>2507 postal officer</td>
<td>4123</td>
<td></td>
<td>Officer, postal (PO) (S)</td>
<td>Bank and post office clerks</td>
</tr>
<tr>
<td>2508 turf accountants clerk</td>
<td>4129</td>
<td></td>
<td>Clerk, accountants, turf</td>
<td>Financial administrative</td>
</tr>
<tr>
<td>2509 recorder in hm dockyard/wages/</td>
<td>4159</td>
<td></td>
<td>Recorder (HM Dockyard)</td>
<td>Other administrative occup</td>
</tr>
<tr>
<td>2510 foundry/iron moulder</td>
<td>5212</td>
<td></td>
<td>Moulder, foundry</td>
<td>Moulders, core makers and</td>
</tr>
<tr>
<td>2511 moulder/moulding steel</td>
<td>5212</td>
<td></td>
<td>Moulder, steel</td>
<td>Moulders, core makers and</td>
</tr>
<tr>
<td>2512 sheet metal worker/skilled</td>
<td>5213</td>
<td></td>
<td>Worker, metal, sheet</td>
<td>Sheet metal workers</td>
</tr>
<tr>
<td>2513 engineer, jig borer</td>
<td>5221</td>
<td></td>
<td>Borer, jig</td>
<td>Metal machining setters and</td>
</tr>
<tr>
<td>2514 moulden tool maker/engineer</td>
<td>5222</td>
<td></td>
<td>Engineer-toolmaker</td>
<td>Tool makers, tool fitters and</td>
</tr>
<tr>
<td>2515 toolmaker in engineering factory</td>
<td>5222</td>
<td></td>
<td>Engineer-toolmaker</td>
<td>Tool makers, tool fitters and</td>
</tr>
<tr>
<td>2516 mechanic/agricultural repairs</td>
<td>5223</td>
<td></td>
<td>Mechanic, agricultural</td>
<td>Metal working production</td>
</tr>
<tr>
<td>Code</td>
<td>Index</td>
<td>Title</td>
<td>Score</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>--------------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>4120</td>
<td></td>
<td>Accountant</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4122</td>
<td></td>
<td>Officer, accounts</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4122</td>
<td></td>
<td>Auditor</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4129</td>
<td></td>
<td>Clerk, counter</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4131</td>
<td></td>
<td>Cataloguer</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4132</td>
<td></td>
<td>Clerk, pensions</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4184</td>
<td></td>
<td>Officer, transport</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4159</td>
<td></td>
<td>Clerk</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4159</td>
<td></td>
<td>Other administrative occ.</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4159</td>
<td></td>
<td>Official, board, water</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4212</td>
<td></td>
<td>Clerk, chief</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>4215</td>
<td></td>
<td>Secretary</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>5111</td>
<td></td>
<td>Farmer</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>5111</td>
<td></td>
<td>Farmers</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>5114</td>
<td></td>
<td>Groundsmen</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>5114</td>
<td></td>
<td>Greenkeepers</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Code</th>
<th>Index</th>
<th>Title</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>9139</td>
<td></td>
<td>Iron and steelworker</td>
<td>30</td>
</tr>
<tr>
<td>9139</td>
<td></td>
<td>Assistant, factory</td>
<td>30</td>
</tr>
<tr>
<td>9139</td>
<td></td>
<td>Worker, factory</td>
<td>30</td>
</tr>
<tr>
<td>9233</td>
<td></td>
<td>Servant</td>
<td>30</td>
</tr>
<tr>
<td>9233</td>
<td></td>
<td>Cleaners and domestics</td>
<td>30</td>
</tr>
<tr>
<td>9230</td>
<td></td>
<td>Docker</td>
<td>30</td>
</tr>
<tr>
<td>9250</td>
<td></td>
<td>Checker, goods</td>
<td>30</td>
</tr>
<tr>
<td>9250</td>
<td></td>
<td>Labourer (docks) (522)</td>
<td>30</td>
</tr>
<tr>
<td>9260</td>
<td></td>
<td>Handler, cargo</td>
<td>30</td>
</tr>
<tr>
<td>9272</td>
<td></td>
<td>Assistant, food and beverage</td>
<td>30</td>
</tr>
<tr>
<td>9273</td>
<td></td>
<td>Water</td>
<td>30</td>
</tr>
<tr>
<td>9279</td>
<td></td>
<td>Worker, hotel</td>
<td>30</td>
</tr>
<tr>
<td>1115</td>
<td></td>
<td>Director (government)</td>
<td>30</td>
</tr>
<tr>
<td>1115</td>
<td></td>
<td>Commissioner (government)</td>
<td>30</td>
</tr>
<tr>
<td>1116</td>
<td></td>
<td>Councillor (local government)</td>
<td>30</td>
</tr>
<tr>
<td>1116</td>
<td></td>
<td>Elected officers and reps</td>
<td>30</td>
</tr>
<tr>
<td>1131</td>
<td></td>
<td>Secretary and company ltd</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Code</th>
<th>Index</th>
<th>Title</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1131</td>
<td></td>
<td>Secretary and company ltd</td>
<td>30</td>
</tr>
</tbody>
</table>
How do I make use of the comparison score?

Tests with large datasets give us an indication of the accuracy of the coding done by Cascot.

**Above 80**: coding takes place automatically

**60 - 80**: reasonably sure that the code shown is the correct one to allocate – but need to look at the alternatives shown with slightly lower comparison scores and check against the additional information you have for coding.

**40 - 60**: there is some ambiguity. Careful consideration of all relevant information is required.

**39 and lower**: Cascot is struggling to find an appropriate match. More information is necessary before coding can be concluded.

Making use of additional information

Click to see input record
**DASISH** funded development of CASCOT v5.0:
- a user interface in 9 selected European languages;
- classification files which permit coding of text in selected languages to the appropriate national occupational classification and to ISCO’08 at four digits;
- a software tool which will facilitate evaluation of coded text files.

Upgraded to facilitate future extension by incorporating additional languages as and when relevant index material becomes available.

**SERISS** funded development of CASCOT v5.5:
- a user interface in 5 additional languages (Arabic, Hindi, Chinese, Indonesian, Russian);
- 5 associated classification files – 31 October 2017
- Performance test reports – 30 November 2017
- Disagreement reports – 30 November 2017
- Revised version of CASCOT software – 12 December 2017
Demonstration of CASCOT v 5.0

English
English with descriptions
Russian
Arabic
Hindi
Chinese
Indonesian

Further information

Email: M.E.Birch@warwick.ac.uk
      R.Day@warwick.ac.uk
      Peter.Elias@warwick.ac.uk

CASCOT
www.warwick.ac.uk/cascot

Institute for Employment Research
University of Warwick
www.warwick.ac.uk/ier
Cascot in practice,
Seriss Survey Experts Workshop
Sue Westerman, 05-09-2017
swtn@cbs.nl

Overview

– Why Cascot: old process versus new process
– Measuring socio-economic variables
– Maintenance and quality checks
Background changing the coding system

Old situation until 2011

By phone (CATI)

Face to face (CAPI)

Computer assisted data collection process

CATI and CAPI questionnaire

Interactive coding module

Redesign social surveys

CAWI / CATI / CAPI: three modes one questionnaire

Coding system suitable for web based interviewing

Shortening of the interview time

IT-policy: no custom-made software applications, only standard tools

New process

Data collection process

CATI CAPI and CAWI questionnaire

Open: occupation, main tasks
Closed: managerial tasks

Open: field of education
Closed: level of education

Open/closed: economic activities

Data processing process

Coding process for occupation

Coding process for education

Coding process for econ. activity

Set of open and closed questions suitable for back-office coding
Collection and processing is the same in all three modes and all surveys

Cascot
User-friendly, easy to make adjustments
Relatively cheap
Suitable for automatic and manual
Suitable for several classifications
Ready to test with Dutch index-file for measuring occupations
Implementing Cascot: education

**Input**
- Field of education: one open question, examples motivate respondents to give precise descriptions.
- Level of education: closed questionnaire.
- Routing open questions on field depend on closed questionnaire.

**Coding**
- Fully automatic coding
- Index: individual names of educational formal and non-formal programmes. In total 31570 entries.

**Quality**
- Non formal education: broad fields with sufficient quality
- Formal education: broad, narrow and detailed fields quality sufficient

Implementing Cascot: occupation

**Data collection**
Open questions on occupation and main tasks
Closed questions to measure managerial tasks

**Coding process**
For measuring ISCO on 4 digit:
- Automatic and manual coding: approx 70% vs 30% resp.
- Index: around 6000 entries
- Automatic coding: occupation, tasks, econ. act., managerial tasks
- Manual coding: in addition level of education, number of employees

For measuring ISCO on 1-3 digits:
- Fully automatic coding possible with indication on quality
- Automatic coding: occupation, tasks
- Manual coding depends on demands on quality and detail
Measuring occupation at 4-digits

**Output**

1. Step 1: Coding based on occupation
   - Score $> 99$
   - Remaining portion score $< 40$ or step3-code

2. Step 2: Coding based on occupation and main tasks
   - Score $> 69$
   - Remaining portion score $< 70$ or step3-code

3. Step 3: Coding based on decision rules using occupation, NACE and managerial tasks
   - Score $< 39$

4. Step 4: Manual coding, using in addition level of education and number of employees
   - Manual coding at all aggregation levels of the classification

**Input**

**ISCO 2008**

- Automatic coding 4-digit level
- Manual coding at all aggregation levels of the classification

---

*Optional manual coding* 15%-20%

Measuring at a higher aggregation level

**Output**

1. Step 1: Coding based on occupation
2. Step 2: Coding based on occupation and main tasks
   - Records with highest score
   - ISCO determined on 1-4 digits
   - ISCO not determined (4%) 15%-20%

**Input**

- 80%
- 20%

New coding process aims at higher flexibility and efficiency
Share of manual coding depends on detail and quality needed
Full automatic coding with indication of quality is possible
Cascot rules: to lower the score, code with unknown or at higher aggregation level.

Detail and quality needed
Other variables coded with Cascot

Economic activity

- Employees
- Self-employed
- Registers
- Questionnaire
- Matching data
- Automatic and manual coding
- Code available from registers

Household Budget survey: products and shops
- Input: two fields that describe product or shop
- Automatic coding process in two steps: record highest score and score 40 or higher → output
- Manual coding: score below 40

Health survey: sports
- Input: one field that describes sport
- Coding automatic and manual

Process maintenance

- Check coded material
- Adjust rules and index in CASCOT
- Test impact on output of adjustments automatic coding
- Decide on implementation

Occupation: two classification files:
1. Automatic coding at 4 digits
2. Manual coding and automatic at higher aggregation

Education, NACE industry, Products/shops, Sports: One classification file

Quality checking: max 10% coded wrong per output category used for publication