Validating occupational coding indexes for use in multi-country surveys

Tijdens, K.G.; Kaandorp, C.S.

Citation for published version (APA):
Tijdens, K. G., & Kaandorp, C. S. (2018). Validating occupational coding indexes for use in multi-country surveys. *Survey Methods - Insights from the Field*. https://doi.org/10.13094/SMIF-2018-00007

Validating occupational coding indexes for use in multi-country surveys

Kea G. Tijdens, Amsterdam Institute for Advanced Labour Studies, University of Amsterdam, Netherlands
Casper S. Kaandorp, Amsterdam Institute for Advanced Labour Studies, University of Amsterdam, Netherlands

16.11.2018

Abstract

Occupational coding in multi-country surveys is mostly a black box: have national survey agencies classified the same occupational titles into the same category across countries? This paper attempts to validate the coding from 5-digit occupational titles into the 4-digit occupational units of the international ISCO-08 classification, based on a comparison of coding indexes from national statistical offices. Two research objectives are central. To what extent are occupational titles in the coding indexes similar, when comparing their English translations? What percentage of similar occupational titles is coded similarly across countries? To answer these questions, we merged titles from 20 coding indexes (18 non-English), resulting in 70,489 records. We translated the titles in English, using online dictionaries and Google translate (4.2% could not be translated). We checked for existent codes of the titles, using ILO's ISCO-08 coding index (10.3% non-existent). The remaining database had 60,559 records, of which 32% had at least one duplicate title (19,044 records). These duplicate records could be aggregated into 5,350 occupational titles. Only 64% of these titles had the same ISCO-08 4-digit code, 70% at 3-digit, 74% at 2 digit, and 80% at 1-digit. Users of multi-country surveys should be cautious when using the 4-digit ISCO-08 codes.

Keywords

isco-08 classification, multi-country surveys, Occupations, validation

Acknowledgement

Copyright

Copyright

Introduction

Occupation is a key variable in socio-economic research. It is used in studies regarding school-to-work transitions, manpower forecasting, the gender pay gap, social stratification, occupational health and safety, processes of professionalization, and alike. Therefore, socio-economic surveys include a question "What is your occupation?", or similar. For capturing the answers to this question typically open text fields are used. These are back-office coded, which is a time-consuming and expensive activity. For web-surveys increasingly survey respondents are asked to identify their occupation from a database of occupational titles. For multi-country surveys this requires a multilingual, multi-country database of titles.

Occupational titles usually are classified in ISCO, which is the worldwide accepted International Standard Classification of Occupations (ISCO), maintained by the International Labour Organisation in Geneva (ILO, 2012). The first ISCO classification dates back to 1958, with updates in 1968, 1988 and 2008 (for details see the [ISCO website](#)). The current version, ISCO-08, is a four-level hierarchical classification with ten major groups at the top of the hierarchy (1-digit), and 436 occupational units at the bottom, the so-called 4-digit units. Long lists of occupational titles, the 5-digit titles, are classified into the 4-digit units. For example the detailed, 5-digit occupational title "Database architect" is classified into the 4-digit occupational unit 2521 "Database Designers and Administrators", that in turn is classified into the 3-digit group 252 "Database and Network Professionals", that in turn is classified into the 2-digit group 25 "Information and communications technology professionals", that finally is classified into the 1-digit major group 2 "Professionals". Most National Statistical Offices (NSOs) have adopted the ISCO-08 classification and have prepared classification schemes, sometimes as the only classification system, sometimes next to an existing national system.

Occupational coding in surveys

The majority of survey respondents do not provide an answer to the 'what is your occupation' question that directly fits into any of the 4-digit occupational unit groups, because these groupes are too aggregated for daily life communication about actual job titles or about occupational titles used in communication beyond the workplace. As most respondents report a title in greater detail, both ISCO-08 and the NSO's have published coding indexes to classify the so-called 5-digit titles into the 4-digit categories. These coding indexes are among others used in the Labour Force Surveys.

In multi-country surveys coding practices mostly rely on national survey agencies and their coding practices are typically based on the national coding indexes, because no multi-language, multi-country coding index exists. This refers to the coding of 5-digit titles into a 4-digit unit, which is challenging task given that a national labour market can have up to 10,000's of different 5-digit occupational titles. Across countries with different languages occupational coding in multi-country surveys is therefore basically a black box, as long as no answer can be given to the issue whether across countries similar 5-digit occupational titles are coded into the same 4-digit category.

A few studies have validated occupational coding practices. Elias and McKnight (2001) conclude that at the aggregate level occupational classifications appear to provide a robust method for the measurement and analysis of skill, but that at 3- and specifically 4-digit level coding discrepancies are noticed. Recoding the parental occupational titles of the SHARE survey showed the same results: at lower levels of aggregation the reliability between coders drops (Belloni *et al*, 2016). Ganzeboom (2014), coding the parental occupations in the European Social Survey, concluded that translating the national batch files of the open-ended answers into English would have been more efficient in terms of resources and would have increased reliability of the codes compared to coding the national batches with the help of national experts.

This paper attempts to validate the coding from 5-digit to 4-digit across countries with multiple languages. This can be done in two ways. First, the original titles in the open text fields of a multi-country survey can be translated and then compared, as suggested by Ganzeboom (2014). Second, the national coding indexes can be translated into English and then compared. In this article we have applied the second approach, because the national coding indexes provide understandable 5-digit titles, which don't have spelling mistakes, weird titles or too aggregated titles. In most countries the national coding indexes are prepared by the national statistical office (NSO), and these indexes assign 4-digit unit groups to lists of 5-digit occupational titles. Typically, NSO's have to manoeuvre between previous and current classifications to ensure comparability over time and between the national classification and ISCO-08.

Translating 5-digit occupational titles are the only way to validate the reliability of their classification into 4-digit occupational units across countries. Here, two viewpoints exist. The first states that occupational titles cannot be translated beyond 4-digits, because occupational boundaries follow country-specific practices and so does the composition of the country's industry and its related occupational structure. Therefore, the classification of 5-digit titles into 4-digit ones cannot be validated by means of cross-country comparisons. Hence, validating the occupational coding in multi-country surveys is not possible, and the black box continues to exist. The second viewpoint states that occupational titles can be translated, thereby allowing for a cross-country validation of the occupational coding. This latter viewpoint departs from the globalising economy and assumes that increasingly across countries similar occupations exist due to a convergence across countries. Global equipment suppliers for example go along with equipment-defined jobs. Globalisation urges the need to understand occupations across countries and similarly the pressure towards cross-country standardization, e.g. QESH auditor. So far, very few studies, if any at all, have explored which viewpoint is true, though it may well be the case that both viewpoints can be true for different groups of occupational titles. For this paper we assume the second viewpoint to be true.

The objective of this article is to validate whether occupational titles listed in national coding indexes are coded similarly across countries. Two research objectives are central in this article:

- To what extent are occupational titles in the coding indexes similar, when comparing their English translations?
- What percentage of similar occupational titles is coded similarly across countries?

Validating multi-country coding indexes requires:

- the availability of coding indexes of NSO's, particularly for non-English speaking countries
- the indexes should apply the ISCO-08 classification and should specify 5-digit occupational titles within the 4-digit unit groups
- the availability of translations of these 5-digit occupational titles into English
- a text mining program that identifies whether the translated English occupational titles are the same, e.g. *office clerk* equals *clerk in the office*.

As will be explained in the next section, for this article we used 20 coding indexes, of which 18 are in another language than English. We merged these coding indexes, translated the initial titles into English, explored to what extent the English titles were similar, and compared if similar titles had the same ISCO-08 4-digit codes.

Data and methods

Data: coding indexes

In the last months of 2015 and the first months of 2016 we collected as many occupational

coding indexes as we could find, as long as these came from countries included in the list of 99 countries, defined in the SERISS project. Departing from the UN's website with lists of national classifications for many countries, we browsed the websites of the National Statistical Offices (NSO) to find their occupational coding indexes (Appendix *List of 99 countries and sources*). For a few countries, where no index was available online, we were able to retrieve a coding index from researchers. Altogether we found indexes for 34 countries. Of these 34 countries, five had another classification than ISCO-08 (Canada, Iceland, India, Italy, Switzerland). These countries either used ISCO-88 or had their own national coding index. Four countries had no coding index beyond the 4-digits of ISCO-08 (Croatia, Malaysia, Norway, Tanzania) and could therefore not be used. Two countries referred to other countries regarding their ISCO-08 coding indexes. Germany referred to the Austrian coding index and Montenegro stated that its index also applied to Serbia. For four countries we encountered technical difficulties. The coding index of Greece was incomplete, as it did not go beyond ISCO-08 code 2122. Translating Hebrew caused too many technical difficulties, among others due to the right-to-left script. Israel therefore could not be included. Due to a technical error the Belgium and the Lithuanian coding indexes had to be dropped. One country, Finland, had an index in two languages (Finnish and Swedish), but because we already had a long Swedish occupational coding index from Statistics Sweden, we did not include the brief Swedish occupational list from Finland. Out of the initial 34 indexes, 19 could thus be merged into the validation database. Finally, we added the English titles of the WageIndicator database of multilingual, multi-country occupational database, that is used in its web survey on work and wages in 92 countries and for which among others the occupational titles in ISCO-08's coding index have been used. For the SERISS project, this database is extended with more occupational titles and more languages (Tijdens, forthcoming 2019). The resulting merged database has 70,489 records of occupational titles from 20 sources in 19 different languages (Table 1).

The number of occupational titles in the database varies largely across the source countries (Table 1). Austria contributes more than 13,000 occupational titles (19% of the titles in the database), Sweden ranks second with more than 12,000 titles (12%), and Albania, Bulgaria, Estonia, Latvia, Netherlands, and WageIndicator contribute each 4,000 to 5,000 titles (6-7%). Czech Republic, Montenegro, Poland, Romania, Slovak Republic, Slovenia, Spain, South Africa, Turkey contribute each 1,000 to 2,500 titles. The coding indexes of three countries (Denmark, Finland, Portugal) comprise less than 1,000 titles. Although ILO's coding index of ISCO-08 should be considered to be the leading index for countries around the world, our inventory of national coding indexes shows that some countries do so much more than others.

Table 1
The merged validation database of translated occupational titles and percentage existent codes by country

#	Coding index	Language	# entries in index	Col. %	% with existent code
1	Albania	Albanian	4273	6.1	75.8%
2	Austria (+ Germany)	German	13395	19.0	100.0%
3	Bulgaria	Bulgarian	5077	7.2	100.0%
4	Czech Republic	Czech	1358	1.9	100.0%
5	Denmark	Danish	564	0.8	97.9%
6	Estonia	Estonian	4715	6.7	100.0%
7	Finland	Finnish	103	0.1	100.0%

8	Latvia	Latvian	4057	5.8	100.0%
9	Montenegro (+ Serbia)	Serbian	2129	3.0	100.0%
10	Netherlands	Dutch	4704	6.7	100.0%
11	Poland	Polish	2443	3.5	94.1%
12	Portugal	Portuguese	708	1.0	92.2%
13	Romania	Romanian	3077	4.4	40.9%
14	Slovak Republic	Slovak	2147	3.0	100.0%
15	Slovenia	Slovenian	2094	3.0	98.3%
16	Spain	Spanish	2502	3.5	31.4%
17	Sweden	Swedish	8617	12.2	72.1%
18	South Africa	English	2252	3.2	97.6%
19	Turkey	Turkish	2214	3.1	100.0%
20	WageIndicator	English	4060	5.8	100.0%
	TOTAL		70489	100%	89.7%

Source: SERISS merged database of coding indexes 2018

Data: check for existent ISCO-08 codes

After merging the occupational titles with their ISCO-08 4-digit codes from the 20 indexes, we checked the existence of the codes by comparing them to the codes in the official ISCO-08 coding index (ILO, 2012). It turned out that 7,275 of the 70,489 titles had codes that did not correspond with any codes in the ISCO classification (10.3%), the so-called non-existent codes. Nine countries have non-existent codes (last column in Table 1). In five countries we find a few, but in four we notice substantial percentages. One reason for the non-existent codes is that NSO's use these codes to enable mapping with their long-existing national classification. Another reason is that NSO's extent a 3-digit with a trailing 0 instead of using the appropriate 4-digit code. For example, the 3-digit code 131 *production managers in agriculture, forestry and fisheries* is extended to the non-existent code1310, instead of using the 4-digit codes 1311 *agricultural and forestry production managers* and 1312 *aquaculture and fisheries production managers.* There may be other reasons too, but we have not asked the NSO's why they use non-existent codes.

The ISC0-08 1-digit groups reveal details about the non-existent codes. Major group 0, the *armed forces occupations*, has the lowest percentage of existent codes, notably 57% of the occupational titles in this group. This is due to three countries. In Albania and Portugal, the army occupations are coded to a greater extent than in the ISCO-08 classification, and in Spain only 76% are coded with existent codes. Take note that Finland, Romania and South-Africa have no occupational titles in this major group. Major group 1, m*anagers*, has the second lowest percentage of existent codes (82%). The remaining groups do not vary to a large extent: group 2 *professionals* (89%), group 3 *technicians and associate professionals* (94%), group 4 *clerical support workers* (85%), group 5 *services and sales workers* (90%), group 6 *skilled agricultural, forestry and fishery workers* (92%), group 7 *craft and related trades workers* (93%), group 8 *plant and machine operators and assemblers* (88%), and group 9 *elementary workers* (89%).

Data: translating titles from coding indexes

The objective of our validation exercise is to explore whether similar 5-digit occupational titles in different countries are coded similarly. Similarity is here defined as a similar English translation of the titles in the national coding indexes. For the concept of 'similar occupations' we fully rely on the occupational title as any information concerning the job content was absent. To explore if similar occupational titles are coded similarly, all non-English titles have been translated into English. Given the huge number of 70,489 occupational titles minus the 6,312 titles from the two English indexes (South Africa and WageIndicator), translation by professional translators was beyond budget. Instead, we developed scripts for using online dictionaries and where these dictionaries did not provide a translation Google translate was used. Where feasible the translations were supplemented with manual improvements. For 2,958 titles (4.2%) the online dictionaries and Google translate did not provide a translation. Three countries had relatively high percentages of non-translatable occupational titles: Austria (12.2%), the Netherlands (7.4%), and Sweden (9.9%). Not surprisingly, the more 5-digit titles in a national coding index, the higher the percentage of non-translatable titles ($r=.80$). The 13,395 occupational titles in the Austrian index were more likely to include some very country-specific and therefore untranslatable titles than the 103 titles in the Finnish index. Translatability did not relate to the number of characters within a single occupational title.

Table 2 provides an example of the translations. Six national coding indexes include a 5-digit occupational title that has been translated into the English occupational title *art historian*. The last column of the table shows the core issue of this paper, notably that three of the six indexes assign the 4-digit code 2633 *philosophers, historians and political scientists* to this occupational title, whereas two others assign the code 2651 *visual artists* and one assigns 2632 *sociologists, anthropologists and related professionals*. This coding dissimilarity will be discussed in the next section.

Table 2
Example of the translation of the national occupational titles into English titles

Coding index	National title	English translation	ISCO-08 4-digit code
Austria	kunsthistoriker	art historian	2633
Estonia	kunstiteadlane	art historian	2651
Latvia	mākslas vēsturnieks	art historian	2633
Netherlands	kunsthistoricus	art historian	2632
Poland	historyk sztuki	art historian	2633
Slovenia	umetnostni zgodovinar	art historian	2651

Source: SERISS merged database of coding indexes 2018

We applied a three-step script-based approach to prepare the English translations for the duplicate titles test. A duplicate title is a title with at least one same title, based on a comparison of the English titles. First, in the English titles we removed dots, non-alphabetic characters, double blanks, quotes, and redundant words as *the*, *of the*, *an*, and alike. Second, we changed plural into singular and female into male occupational titles. Third, we moved words to other parts of the sentence if that identified duplicates. The results will be discussed in the next section. The full database is attached to this article (Appendix *SERISS-Deliverable 8-4 coding indexes 20180702*). It has 70,489 records and has variables indicating the initial and translated labels and their ISCO-08 4-digit code, variables indicating the country, the ISCO-08 1 to 3-digit codes, and binary variables indicating whether the record was translatable, had an existing code and had a duplicate title.

Results

To what extent are occupational titles in the coding indexes similar?

Our first research objective addresses the extent to which translated occupational titles are similar. We exclude the non-translated occupational titles and those with non-existent codes, resulting in a database with 60,559 records. Records with English titles that occur at least twice were assigned a duplicate score of 1 whereas records without duplicates are assigned value 0. As table 3 shows, 31% of the records have English titles that are at least twice present in the database. We use the words *duplicate titles* for these records. This 31% is lower than we initially hoped.

Table 3
Frequency and percentage of single and duplicate occupational titles

duplicate_label_eng	Frequency	Percent
Single title	41515	68.6
Duplicate title	19044	31.4
Total records	60559	100.0

Source: SERISS merged database of coding indexes 2018, selection: records with translated titles and existent codes

Several factors may explain why the percentage of single titles is high. The first one is that the percentage of duplicate translations varies largely across countries, ranging from 4% of the English translations of the Danish titles to 68% of the translations of the Turkish occupational titles and 83% of the South-African English titles (see Table 4). Most likely, some countries adapt the ISCO-08 coding index to a much larger extent than other countries do, which affects the share of duplicate titles.

Table 4
Percentage duplicate titles by country

country	Share duplicate	N	Std. Dev.
Albania	15.4%	3163	0.361
Austria (+ Germany)	32.3%	11946	0.468
Bulgaria	16.6%	5077	0.372
Czech Republic	10.5%	1358	0.307
Denmark	4.3%	540	0.202
Estonia	30.7%	4702	0.461
Finland	20.4%	103	0.405
Latvia	24.7%	3998	0.431
Montenegro (+ Serbia)	38.1%	2085	0.486

Netherlands	31.3%	4481	0.464
Poland	23.2%	2213	0.422
Portugal	26.6%	651	0.442
Romania	16.3%	1222	0.369
Slovak Republic	16.0%	2126	0.367
Slovenia	34.6%	2036	0.476
Spain	37.8%	783	0.485
Sweden	28.2%	5609	0.450
South Africa	83.2%	2197	0.374
Turkey	67.2%	2210	0.469
WageIndicator	47.2%	4059	0.499
TOTAL	31.4%	60559	0.464

Source: SERISS merged database of coding indexes 2018, selection: records with translated titles and existent codes

A second explanation relates to the fact that some languages use brief sentences instead of one or two words to identify an occupational title. For example, the Bulgarian occupation *"Работник, ремонт на стоматологични инструменти"* is translated into *worker repairing dental instruments.* In the same way the Serbian title *"Radnik na proizvodnji kompota"* is translated as *a worker on the production of compote.* These sentence-like occupational titles hamper the identification of duplicates. We assume that the number of characters in the initial occupational title affects the incidence of duplicates. Indeed, the titles for which a duplicate is identified have on average 17.6 characters, whereas the titles without a duplicate count on average 28.8 characters. When broken down by country, the data shows that the occupational titles in South-Africa have on average the lowest number of characters (19.6), whereas the Czech, Danish, Slovak and Portuguese indexes have the largest number of characters (on average more than 35). These differences reflect different name-giving concepts of occupational titles across countries. Table 5 provides an insight how the number of characters in a national title and linguistic styles affect the incidence of duplicate titles, using a regression model. The results show that this effect holds, also when controlled for ISCO-08 1-digit groups, as the odds ratio decreases 9%.

Table 5
Logistic regression for the binary variable duplicate (1=duplicate present, 0 = no duplicate present) by length of initial wording and by ISCO-08 1-digit categories (Ref: armed forces (group 0), and elementary (9) and agricultural occupations (6)), N= 60,559

Variables	Model 1	Model 2
	Exp(B)	Exp(B)
Length: number of characters in national title (1 to 211)	0.913	.909
ISCO group 1 Managers		2.083

ISCO group 2 Professionals		2.029
ISCO group 3 Technicians and associate professionals		1.621
ISCO group 4 Clerical support workers		1.118
ISCO group 5 Service and sales workers		1.252
ISCO group 7 Craft and related trades workers		.827
ISCO group 8 Plant and machine operators and assemblers		.876
Constant	3.411	2.747
Chisq,(df)	10631.74, (1)	12059.35,(8)

Source: SERISS merged database of coding indexes 2018, selection: records with translated titles and existent codes

A third explanation relates to the large variation of the number of 5-digit occupational titles within the ISCO-08 4-digit units across the indexes. Within occupational unit 9123 *window cleaners* the twenty countries provide between zero and two 5-digit occupational titles, and as a consequence the number of duplicate translations is high in this 4-digit occupational unit. In contrast, within occupational unit 8160 *food and related products machine operators* the twenty countries provide between 0 and 133 5-digit occupational titles, ranking Austria with 133 and Spain with 122 titles on top. With 211 titles in the 4-digit occupational unit 5223 *shop sales assistants* Austria has the highest number of 5-digit occupational titles within a single ISCO-08 4-digit unit. The chance of finding duplicates is therefore much smaller in these units. Table 6 confirms this expectation.

Table 6
Logistic regression for the binary variable duplicate (1=duplicate present, 0 = no duplicate present) by number of titles per index in each 4-digit category and by ISCO-08 1-digit categories (Ref: armed forces (group 0), and elementary (9) and agricultural occupations (6)), N= 60,559

Variables	Model 1	Model 2
	Exp(B)	Exp(B)
Length: number of characters in national title (1 to 211)	.991	.991
ISCO group 1 Managers		1.104
ISCO group 2 Professionals		1.834
ISCO group 3 Technicians and associate professionals		1.278
ISCO group 4 Clerical support workers		1.054
ISCO group 5 Service and sales workers		1.446
ISCO group 7 Craft and related trades workers		.981

ISCO group 8 Plant and machine operators and assemblers		.683
Constant	0.572	.466
Chisq,(df)	741.66, (1)	1935.14,(8)

Source: SERISS merged database of coding indexes 2018, selection: records with translated titles and existent codes

In conclusion, for 69% of the titles no duplicate occupational title can be identified in the merged database of 60,559 records. The results show that the incidence of a duplicate title is affected by both the number of characters in the initial occupational title and the number of titles within one 4-digit group per index. The odds ratio for duplicate title decreases when the initial titles are longer and when there are more titles within one group.

Similar occupational title, not similar code

The second objective of our study is to explore what percentage of duplicate occupational titles is coded similarly across countries, thus have been assigned the same ISCO-08 4-digit code in the national coding indexes. This analysis is limited to the 19,044 records with duplicate occupational titles in the database.

Let's start with an example. Table 7 shows the classification for a selection of five 5-digit occupations in the database. Two coding indexes include the occupation *wine consultant*, and both are assigned the same ISCO-08 4-digit code, thus the standard deviation is 0 for this occupational title. Seven coding indexes include the occupation *wine grower*, of which six are assigned code 6112 *tree and shrub crop growers* and the other is coded 6111 *field crop and vegetable growers*, resulting in a standard deviation (s.d.) of .38. Five coding indexes include the occupation *wine maker*, and these are assigned four different codes: 2145 *chemical engineers*, 6112 *tree and shrub crop growers*, 7514 *fruit, vegetable and related preservers*, and 7515 *food and beverage tasters and graders*, resulting in a s.d. of 2202.33. The two occupations *wine taster* and *wine waiter* are each assigned the same codes, and therefore the s.d. is 0.

Table 7
Five occupational titles, the standard deviation across their ISCO-08 4-digit codes in the national indexes, and the number of indexes where the occupational title is present

Occupational title in English	S.d. of 4-digit across indexes	# coding indexes
Wine consultant	0	2
Wine grower	0.38	7
Wine maker	2202.33	5
Wine taster	0	5
Wine waiter	0	3

Source: SERISS merged database of coding indexes 2018, selection: records with translated titles and existent codes of five occupational titles

As said, the merged database has 19,044 records with occupational titles that have at least one duplicate English title and of these records only 54% has the same 4-digit code as all their duplicates. To explore the coding of duplicate titles further, we aggregate the 19,044 records into the occupational titles, resulting in a dataset with 5,350 titles, hence 3.6 records per title. Of these 5,350 titles, 64% have a standard deviation of 0, hence all records have the same ISCO-08 4-digit occupational code (Table 8). The similarity improves when comparing the 3-digit codes of the occupational titles. The number of titles with s.d.=0 increases to 3,754 (70%). The similarity improves further when comparing the 2-digit codes. The number of titles increases to 3,968 (74%). And finally, when comparing the codes at 1-digit level, the similarity increases to 80%.

Table 8
Similarity at 4-, 3-, 2- and 1-digit level of the 5,350 records with duplicates

Similarity at digit level	N	% of 5,350
Similar 4-digit code	3412	63.8
Similar 3-digit code	3754	70.2
Similar 2-digit code	3968	74.2
Similar 1-digit code	4291	80.2

Source: SERISS merged database of coding indexes 2018, selection: aggregated data from records with translated titles, existent codes, and duplicate titles

Conclusion and discussion

Occupational coding in multi-country surveys is mostly a black box: have national survey agencies classified the same occupational titles into the same category across countries? By merging occupational titles and their ISCO-08 4-digit codes from 20 national coding indexes from national statistical offices, this paper attempts to validate the classification of 5-digit occupational titles into the 4-digit occupational units of the international ISCO-08 classification. In the merged database of 70,489 records we find that 4.2% records could not be translated, and 10.3% had non-existent codes. In the remaining database with 60,559 records, we find that only 31% has at least one duplicate title. Of these 19,044 records, we find that 54% of the titles have similar codes as all their duplicates. When aggregating the duplicate records into a database of occupational titles (5,350 records), we find that 64% have the same 4-digit code across the indexes, 70% has so when limiting the comparison to 3-digit codes, 74% has so for 2-digit codes, and 80% has same codes when comparing the 1-digit code.

What can be learned from these findings? First, users of multi-country survey datasets should be cautious when using the 4-digit ISCO-08 codes, and rather restrict to higher level digits. These recommendations are in line with those of the earlier mentioned validation studies (Elias and McKnight, 2001; Belloni *et al*, 2016). Second, translating lists of 5-digit occupational titles from a non-English language to English cannot fully rely on online dictionaries, as was shown by the 4% of non-translatable titles in our merged database. This needs to be supplemented with human translators. Third, the limitations of survey datasets coded at national level for comparative study also serves to emphasise the value of the work being carried out under the SERISS project to build a harmonised cross-national database of occupations. Such a database can be used for self-identification or for interviewer-based identification based on a search tree or on semantic matching in multi-country surveys (Tijdens 2015, Tijdens 2019 forthcoming). Even though such a database may not fully capture the occupational composition of the national labour market, it has no inconsistent occupational

coding across countries, as identified above.

Appendices

List of 99 countries and sources

SERISS-Deliverable 8-4 coding indexes 20180702

References

1. Belloni, M., Brugiavini, A., Meschi, E., Tijdens, K.G. (2016) Measurement error in occupational coding: an analysis on SHARE data, *Journal of Official Statistics*, 32 (4): 917–945.
2. Elias, P., McKnight, A. (2001) Skill measurement in official statistics: recent developments in the UK and the rest of Europe, *Oxford Economic Papers* 3: 508-540.
3. European Commission, Directorate-General for Research & Innovation, Research infrastructure (2015). *ANNEX 1 (part A) RESEARCH AND INNOVATION ACTION NUMBER — 654221— SERISS*. Brussels, European Commission.
4. Ganzeboom HBG (2014) Coding and scaling of parental occupations in the European Social Survey. University of Amsterdam. Presentation given at Amsterdam, InGRID Workshop, February 10 2014
5. ILO (2012) *International Standard Classification of Occupations Isco-08. Volume 1 Structure, Group Definitions and Correspondence Tables*. Geneva: International Labour Office.
6. Tijdens, K.G. (2015) Self-identification of occupation in web surveys: requirements for search trees and look-up tables. *Survey Methods: Insights from the Field*, DOI:10.13094/SMIF- 2015-00008.
7. Tijdens, K.G. (forthcoming in 2019) *Database of Occupations for 99 countries*. Deliverable D8.3 of the SERISS Project funded under the European Union's Horizon 2020 Research and Innovation Programme Ga No: 654221. Available at (forthcoming): www.seriss.eu/resources/deliverables .