Do's and dont's of occupation coding

Harry B.G. Ganzeboom
PIAAC Conference
Bologna, January 20 2010

Question format -- open

- Because occupations are complicated, it is often advised to collect the information in an open format.
- Underlying assumption is that no set of closed questions can sufficiently measure the required details.
- Questions usually have two elements:
  - Job title
  - Describe major duties and task
- This information is recorded verbatim and then post-processed (coded in the office) using a standard classification of occupations, such as ISCO-08.

Do's and Dont's

- I have produced a cookbook on occupation coding that covers:
  1. Choosing the occupational classification
  2. Creating a coding file
  3. Semi-automatic and automatic coding
  4. Selecting and training coders
  5. Multiple coders and dividing up the work
  6. Coding
  7. Quality checks
  8. Archiving
- The document is written for ISCO-88, but I have added and extension on ISCO-08
- See: http://home.fsw.vu.nl/hbg.ganzeboom/pdf/

Common problems of occupation coding

- Recording open information is already a lot of work.
- Coding occupations very often is the major part of post-processing survey information. Occupation coding is almost always late (or even never completed).
- Coders are hard to monitor.
- You always end up with a certain amount of vague and uninterpretable information.

Two major problems

- How to speed up?
- How to find the right code?
- And how do we know?

- The DO's and DON'T's should answer these questions…

ADVICE 1 + 2

1. The most common source of confusion (respondents and interviewers) is between industry (firm) and occupation (job). The best way to avoid this is to ask for both in the following order:
   - What does your firm do or produce?
   - What do you do?
2. The confusion still arises: it is therefore useful to do occupation and industry coding from the same information (coding file).

ADVICE 3

- Always transfer answers to open questions in electronic format (strings). Never code information questionnaire-by-questionnaire.
- Transferring this information is rather low-level clerical work.
- If you use Excel, be aware of the dangers of its capacities to self-complete strings.

Coding file

- To code occupation, it is useful to collect the occupational information in a coding file. This contains at a minimum:
  - ID
  - Variable name
  - Strings for job title, duties-tasks
- Additional information can be (if asked in interview):
  - Status in employment
  - Supervising status
  - Industry / firm
  - Firm / farm size
  - Required qualifications.

What should NOT be in the coding file?

- Coding file should NOT contain:
  - Education
  - Earnings
  - Age
  - Gender
- Coders should NOT be allowed to 'peek' at these non-occupational characteristics. This is another reason why coding should not be done in questionnaires.

Multiple occupations

- ADVICE 4: If multiple occupations are asked (respondent, spouse, father, mother, careers), all information should be collected in one coding file in LONG format.
- Having access to multiple occupations is extremely helpful to assess quality of coding. I wil discuss later how.

National classifications

- Many countries have developed and use their own national classifications.
- Some are developed by research agencies, but more often by the government statistical agency. They are often revised with 10-year (census) interval.
- If they exist, they are likely to come with a manual and other materials in the national language. This is very useful.
- Over recent years, there has been a strong move to adopt the International Standard Classification of Occupations at a national tool (sometime slightly adapted).
- If you use a national classification to code (and then convert), you need to be prepared to build in a stage in which you evaluate and correct the conversion.

Using old work

- If you have a database of occupation already coded (in ISCO or some other classification), it is extremely useful to match the new coding file with the existing coding file.
- This is a major time saver.
- However, the codes may be wrong. Be prepared to revise everything (and let the coders do this independently of one another).

ISCO

- The stated goal in ISCO-08/ISCO-88 is to organize the information primarily by skill level. The order of the major groups is supposed to be according to the levels of the International Standard Classification of Education:
  - Tertiary
  - Higher / Post-Secondary
  - Lower Secondary
  - Primary
- However, even the Introduction (ISCO-88 manual) shows that this is not (consistently) applied.

ISCO-08 Major Groups

1000 Legislators, Senior Officials and Managers
2000 Professionals
3000 Technicians and Associate Professionals
4000 Clerks
5000 Service and Sales Workers
6000 Skilled Agricultural and Fishery Workers
7000 Craft and Related Trades Workers
8000 Plant and Machinery Operators
9000 Elementary Occupations

Please note ...

- Unlike the ISCO manual, I write the codes of these groups with trailing 000. ADVICE 5: Follow this good idea.
- This is a very useful habit, and fortunately ISCO-08 allows this (this was not true in ISCO-58 and ISCO-68).
- Some titles have been slightly abbreviated.
- The ordering of groups is not fully consistent with skill level. This is in particular true for (1000) managers and (5200) Sales Workers. Implicit organization by authority and manual/non-manual.

Major, sub-major, minor, unit

- 1000 Managers
  - 1100 Chief Executives, Legislators, Sen. Officicials
    - 1110 Legislators and Senior Officials
      - 1111 Legislators
      - 1112 Senior Government Officials
      - .....
    - 1120 Managing Directors and Chief Executives
  - 1200 Administrative and Commercial Managers
    - 1210 Directors and CEO's
    - 1220 Production and Operations Department Mang.
    - 1230 Other Department Managers
  - 1300 Production Managers
  - 1400 Hospitality, Retail and Other Services Managers

The use of the hierarchy by coders

- For accurate measurement, it is much more important to get the Major and Sub-Major groups (first two digits) right than the last two digits.
- ADVICE 6: First code the first two digits.
- For experienced coders, this can be done without consulting the manual (provided that they are willing and able to correct their initial choices).
- This is an important time-saver.
- ADVICE 7: train your coders primarily to understand the differences between the 10 major groups (and the logic behind it)!!

Ambiguities with the major groups

- What is an (associate) professional?
- Where to put farmers and farm workers?
- Shop owners, work supervisors and foreman.
- What is the difference between a craft worker and a machine operator?

Unfortunately, these questions do not have a satisfactory and conclusive answer.

Associate / Professionals

- Professionals are highly skilled workers with expert knowledge (and skills) usually obtained in higher tertiary (university) training.
- Associate professionals are highly skilled workers with expert knowledge (and skills) usually obtained in lower tertiary training. They are associate to (full) professionals.
- Odd exception: Doctors (2210) and Nurses (2220) are both (full) Professionals.

Farmers (in ISCO-98)

- 1301 Production Manager Agriculture
- 6000 Skilled Agricultural Workers
- 6100 Market-Oriented Skilled Agr. Wrk.
- 6300 Subsistence Farmers
- 9200 Agricultural Labourers

In particular the choice between 6100 and 1311 may cause confusion. I prefer to avoid 1310. According to the definitions it is to be restricted yo managers of very large farms.

Shop-owners, supervisors, foremen

- ISCO-08 avoids all reference to self-employment with one exception. Shop-owners are to be classified as 5221 (Shop Keepers).
- Supervisors and Foremen can be coded as:
  - 1300 Production Managers
  - 3120 Production Supervisors
  - 3341 Office Supervisors
  - 5150 Building Supervisors
  - 5222 Shop Supervisors (!)

Craft/machine workers

- A whole list of occupations duplicates between 7000 (Craft Workers) and 8000 (Machine Workers), e.g.:
  - 7510 Food Processing Workers
  - 8160 Food and related Product Machine Opr
- I tend to prefer the 7000 versions using the 'skill level rule'.

Rules for solving ambiguous cases

- Often job descriptions are ambiguous because they contain multiple tasks. Rules to resolve these problems are in the Introduction of the ISCO-88 manual.
- To be applied in this order:
  - ***Majority rule***: if one task prevails (takes a majority of the time), choose this code.
  - ***Skill level rule***: if a description contains tasks of different levels, give preference to the highest level.
  - ***Production rule***: if a description contains production and sales task, give preference to coding by production.

How to process crude information

- Often respondents do not provide information enough to warrant detailed four-digit coding.
- Using one or two digits is often a good solution:
  - Skilled Worker: 7000
  - Semi-skilled Worker: 8000
  - Foreman 3120
  - Manager 1000 or 1300 (??)
- Occasionally ISCO provides n.e.c. (not elsewhere classified) categories.
- Mixing up 1- 2- 3- and 4-digit coding is not a problem, as long as you use trailing zeroes.

Do we need 3- or 4-digit coding?

- To many users 3- or 4-digit coding seems overly detailed and laborious. Do we really need this information?
- For sociological purposes (using the socio-economic status of occupations), 2.5 digit is enough. I.e. 2-digit codes pick up most of the relevant distinctions, but note e.g.:
  - 1200 and 1300 contain Farm Managers
  - 2200 contains Doctors and Nurses.
  - 2300 contains Primary Teachers and University Professors.

It is not a lot of work to code the last two digits

- Projects often settle for coding only the first digit or first two digits.
- This does NOT save half of the work.
- If you sort the coding file by the first two digits, adding in the final two digits is not a lot of work, but this time you need to use the manual!
- This detailed round is in fact very useful in reviewing the choices that have initially been made.
- ADVICE 8: Code all four digits, but in two rounds.

Bad coding practices

- Coding is done by a single, 'expert' coder.
- Coders are trained by doing the job.
- Coders do not have access to manuals.
- If multiple coders are employed, they consult each other about difficult cases.

Good coding practices

- ADVICE 9: Employ multiple coders.
- ADVICE 10: Coders should be trained and instructed, NOT corrected.
- ADVICE 11: Coders should not communicate to one another, but work independently.
- ADVICE 12: Coders should have access to the full classification and in particular to the (English language) manual.
- ADVICE 13: The best coding is (independent!) double coding.

When distributing the coding file over coders

- ADVICE 14: Give the coders each a random part to code. So:
  - Do not give one A..M and the other N..Z.
  - Do not give one the father's and the other the respondents' descriptions.
- ADVICE 15: Make sure that you have all the information before you start. Adding in late interviews usually is a lot of trouble and blurs the coding design.

The coding process

- Coders should first code the first two digits (without much consultation of the manual), and then the last two digits (with much consultation of manual).
- Coders should be able to sort the coding file (without damaging it).
- Coders should have access to the full documentation and to the internet.

Double coding

- Double coding is an expensive, but invaluable way to improve coding quality:
  - If you can operate multiple indicator models in your analysis, have all occupations double coded and maintain the codes with the data files
  - If your only purpose is to assess the quality of coders, have their coding tasks partly overlap. Even as little as 10% overlap of a large task helps quite a bit.

Recruiting, instructing and monitoring coders

- Not everybody likes occupation coding.
- Dividing up the work over multiple coders and having the task done quickly, makes it more fun.
- Instruction should concentrate on the logic of the classification, not on the coding files. Emphasize the major groups.
- Review and instruct. Do NOT correct (but leave the corrections to the coders)!

Evaluating coder quality

- An obvious way to look at coder quality is to have part of the coding file overlap between coders.
- However: if you have only two coders per occupation, you get no information who is the best.
- However, if you have multiple occupations to code (e.g. respondent / father), you can learn about coder quality by comparing between-occupation correlations between coders, even when you have only two coders.
- The DO's and DONT'S document shows a MTMM (lisrel) model how to evaluate this information.