Using Self-Generated ID-Codes to Re-identify Young People in Panel Studies

Challenges and Solutions

ESRA Conference 2019 - Surveying Children and Young People 3

Speaker: Dipl.-Soz. Robert Lipp
Research Centre of Demographic Change (FZDW)
Frankfurt University of Applied Sciences
Nibelungenplatz 1
60318 Frankfurt/Main

Tel.: +49 69 1533-3821
E-mail: robert.lipp@fzdw.de

Research Centre of Demographic Change
www.fzdw.de

The GUS study

- Health Behavior and Injuries in School Age (GUS)
- Main objective
  - Identifying potential causes for injuries of pupils occurring at school
- Funding by the German National Accident Insurance (DGUV)
- Nation-wide, panel survey of about 10,000 students (injured and not injured)
  - 600 classes
  - 150 schools
- First wave started in school year 2014/15 (5th grade)
  - Overall six waves till school year 2019/20 (10th grade)
- Field work is done by ourselves

Methods

- Stratified random sample of German students in secondary education
- Annual survey of the same students

- Combined CAPI/CASI
  - Data collection from the whole grade of the school
  - Interviewer visits school and introduces the survey
  - Students answer the questionnaire themselves using a tablet computer

- Self-generated ID-codes
  - Re-identify students across the panel waves
  - Ensure anonymity of the students

Self-generated IDs

- Four elements
  - First letter of own first name
  - First letter of mother's first name
  - First letter of father's first name
  - Month of birth

Data structure (wide format)

School-ID	Student-ID Wave 1	Student-ID Wave 2	Student-ID Wave 3	Student-ID Wave 4
22720	dmp09	dmp09	dmp09	dmp09
22720	fmt10	-	-	-
22720	-	fst10	-	-
22720	jmw04	jmw04	jmw04	jmw04
22720	met03	-	-	met03
22720	-	-	me003	-
22720	-	njd10	njd10	-
22720	nun04	nun04	-	-
22720	-	-	nuk04	-
22720	uet06	uet06	uet06	uet06
22720	…	…	…	…

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Types of matches

- Correct positive match
  - Same person in both waves, same code in both waves
- Correct negative match
  - Different persons in the two waves, different codes in the two waves
- False positive match
  - Different persons in the two waves, same code in both waves
  - May result from coincidental matching of the code elements
- False negative match
  - Same person in both waves, different codes in the two waves
  - May result from memory problems or typing errors

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Dealing with false matches

- False **positive** matches:
  - Refine code elements
  - Increase length of code
  - Use time-constant variables for verification

- False **negative** matches:
  - Record linkage!

Record linkage

- Fuzzy string-merge

- Stata Ado “Reclink”
  - Bigram algorithm
  - Adjustable matching score (we used 0.86)
  - Allows for control variables (we used gender, year of birth, and number of older siblings)

- Only two waves can be matched at a time
  - ➢ To link four waves, a total of six record linkages is needed

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Data before record linkage

School-ID	Student-ID Wave 1	Student-ID Wave 2	Student-ID Wave 3	Student-ID Wave 4
22720	dmp09	dmp09	dmp09	dmp09
22720	fmt10	-	-	-
22720	-	fst10	-	-
22720	jmw04	jmw04	jmw04	jmw04
22720	met03	-	-	met03
22720	-	-	me003	-
22720	-	njd10	njd10	-
22720	nun04	nun04	-	-
22720	-	-	nuk04	-
22720	uet06	uet06	uet06	uet06
22720	…	…	…	…

Data after record linkage

Matched ID	School-ID	Student-ID Wave 1	Student-ID Wave 2	Student-ID Wave 3	Student-ID Wave 4
dmp09	22720	dmp09	dmp09	dmp09	dmp09
fst10	22720	fmt10	fst10	-	-
jmw04	22720	jmw04	jmw04	jmw04	jmw04
met03	22720	met03	-	me003	met03
njd10	22720	-	njd10	njd10	-
nuk04	22720	nun04	nun04	nuk04	-
uet06	22720	uet06	uet06	uet06	uet06
…	22720	…	…	…	…

Results

- **Confirmation variable:** “Have you participated in this survey in the last year?”

- **Without record linkage:**
  - 14,652 of 36,823 cases (40%) linked across all four panel waves (3,663 students)
  - 99% correct negative matches according to confirmation variable in wave 2
  - 89% correct positive matches

- **With record linkage:**
  - 15,872 of 36,823 cases (43%) linked across all four panel waves (3,968 students)
  - 98% correct negative matches according to confirmation variable in wave 2
  - 93% correct positive matches

- In total 7% of the overall cases could be matched through record linkage

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Summary

- Self-generated codes work well in the school environment
  - Small units of ~100 pupils per school
- Larger sample units may suffer from a higher rates of false positives
  - ➢ Possible solution: Longer code
- Record linkage can improve matching rates without drawbacks to data quality
  - ➢ Control variable can help with the assessment
- However: Gets complicated when linking many waves
- Manual review of the linked cases recommended

Final remark

Especially in panel studies, every case counts!