Introduction to EU-SILC from a research perspective

Heike Wirth
GESIS – Leibniz Institut für Sozialwissenschaften

DwB-Training Course on EU-SILC, February 13-15, 2013
Romanian Social Data Archive at the Department of Sociology
University of Bucharest, Romania
Overview

Strengths of EU-SILC data
Perfect data doesn’t exist
Finally ...
Strengths of EU-SILC

• Coverage of countries (all EU-members & some other countries)

• Coverage of topics
  • 5 core areas: basic data; income; social exclusion; labour market participation; housing

• Sufficiently large sample sizes
  • Allowing for analysis of small groups
  • Allowing for analysis of small regions
  • Leading to higher precision
Strengths of EU-SILC

- Household survey (all members >= 16 of household are included, basic information for members < 16)
  - Unit of analyses: Household, family, couples, individuals etc.

- provision of both cross-section & longitudinal microdata at the level of persons & households

- Data access for research purposes (timeliness, no fees)
Strengths of EU-SILC

• Common framework for all participating countries regarding the implementing of measures:
  • harmonized list of target variables to be transmitted to Eurostat;
  • common guidelines and procedures;
  • common concepts and classifications

• High flexibility within this common framework is an essential feature of EU-SILC (accommodate the needs of NSIs)
Strengths of EU-SILC

All in all

- EU-SILC is a major data source of comparative analysis for income, material deprivation, poverty and living conditions in Europe
- EU-SILC is the standard source for social reporting in Europe

thus comparability, reliability and validity of the data is crucial, no matter whether the data is used
  - for policy making and administrative purposes at the European level or
  - for cross-national research purposes
Perfect data doesn’t exist – Main issues of concern

• Cross-national comparability of EU-SILC?
  • Flexibility in modes of data collection
  • Flexibility in wording of questionnaires (output harmonisation)
  • Flexibility in sample design

• Different types of reference periods

• Need for more information in the Userdata Base
  • Methodological information
  • Substantial information
  • Link of cross sectional & longitudinal data
Flexibility in modes of data collection

• Sources of data could be:
  • survey(s)
  • register(s)
  • combination of survey(s) & register(s)

➢ Data could come from one source (survey or register) or two sources (survey & register)

➢ Issue of concern: Measurement in surveys and registers may be based on different concepts, e.g.
  ➢ Income in register tax-based (what about non-tax income?)
  ➢ Records of earnings might refer to different points in time
  ➢ Information about employment status might differ
Flexibility in modes of data collection

(99.9%) Information/Interview completed from

Survey 73.1%

Register 3.4%

Both: Survey & Register 22.8%

Full Record Imputation 0.7%

Survey countries
BE, BG, CY, CZ, DE, GR, ES, EE, FR, HU, IT, LT, LU, MT; AT, PL, PT, RO, SK, UK

Register countries
Danmark, Suomi, Sverige, Nederland, Iceland, Latvia, Norway, Slovenia, Ireland

Source: UDB_c10R_ver 2010-2 from 01-08-12; own calculation
Moreover there are mixed modes of data collection in surveys

- CATI (Computer Assisted Telephone Interview)
- CAPI (Computer Assisted Personal Interview)
- PAPI (Paper and Pencil Personal Interview)
- self-administered (respondent completes the questionnaire him/herself)
- proxy-interviews (respondent has someone else answer the questions for him/her)
Flexibility in modes of data collection

SURVEY COUNTRIES

Interview
73.1%

Face to Face
• Capi 29.1%
• Papi 42.3%

Cati
3.9%

Self-administered
5.5%

Proxy Interview
18.3%

REGISTER COUNTRIES

Both: Interview & Register
22.8%

Face to Face
• Capi 0.4%
• Papi 11.7%

Cati
65.6%

Self-administered
0.4%

Proxy Interview
22.0%

Source: UDB_c10R_ver 2010-2 from 01-08-12
Flexibility in modes of data collection

- Substantial findings based on EU-SILC could also be affected by type of interview and thus reduce the comparability between countries

- E.g. quality of proxy interviews might depend on the reason of the proxy interview
  1. a respondent is not willing or able to give interview
  2. producers take proxy interviews as a mean to lower costs

- in the first case the proxy interview might be cofounded with other characteristics like age or sex, thus the proxy effect is not distributed at random => impact on distributions

- in the second case the data provider might make efforts for a random selection of proxy respondents thus the proxy effect should less likely confounded by other characteristics
Share of proxy interview by country – ‘register countries’

<table>
<thead>
<tr>
<th>Country</th>
<th>% of proxy interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceland</td>
<td>0.0</td>
</tr>
<tr>
<td>Nederland</td>
<td>1.2</td>
</tr>
<tr>
<td>Sverige</td>
<td>2.1</td>
</tr>
<tr>
<td>Ireland</td>
<td>23.7</td>
</tr>
<tr>
<td>Latvia</td>
<td>23.7</td>
</tr>
<tr>
<td>Norway</td>
<td>23.8</td>
</tr>
<tr>
<td>Slovenia</td>
<td>24.6</td>
</tr>
<tr>
<td>Suomi</td>
<td>42.7</td>
</tr>
<tr>
<td>Danmark</td>
<td>48.6</td>
</tr>
</tbody>
</table>

As a rule only 1 person in hh is interviewed, who answers also for all other hh members

Datasource: UDB_c10R_ver 2010-2 from 01-08-12, own computation
Share of proxy interviews by country – ‘survey countries’

Survey countries: % of proxy interviews
(Data source: UDB_c10R_ver 2010-2 from 01-08-12)

<table>
<thead>
<tr>
<th>Country</th>
<th>% Proxy Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovak Republic</td>
<td>4.3</td>
</tr>
<tr>
<td>Ellada</td>
<td>8.0</td>
</tr>
<tr>
<td>Belgique</td>
<td>8.6</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>10.7</td>
</tr>
<tr>
<td>Oesterreich</td>
<td>13.7</td>
</tr>
<tr>
<td>Romania</td>
<td>15.3</td>
</tr>
<tr>
<td>Lithuania</td>
<td>15.6</td>
</tr>
<tr>
<td>Deutschland</td>
<td>18.8</td>
</tr>
<tr>
<td>Italia</td>
<td>19.0</td>
</tr>
<tr>
<td>Poland</td>
<td>19.2</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>19.3</td>
</tr>
<tr>
<td>Hungary</td>
<td>19.8</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>20.1</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>20.3</td>
</tr>
<tr>
<td>Portugal</td>
<td>20.5</td>
</tr>
<tr>
<td>Espana</td>
<td>21.9</td>
</tr>
<tr>
<td>Cyprus</td>
<td>23.0</td>
</tr>
<tr>
<td>Estonia</td>
<td>24.3</td>
</tr>
<tr>
<td>France</td>
<td>27.6</td>
</tr>
<tr>
<td>Malta</td>
<td>28.9</td>
</tr>
</tbody>
</table>
Research Example – Lohmann (2011)


Research interest
• do findings differ in a systematical way depending on whether the information stems from one data source or from different data sources?

Assumption:
• Consistency of individual income and employment depends on the approach of data collection
Operationalisation

• **Consistency** is measured by the overlap of employment and income of employment

• Two measures to define working & non-working population
  1. based on *activity status* (>= 1 months employed)
  2. based on *income* (>= 1 euros earnings)

• Four approaches of data collection
  • Survey / personal interview
  • Survey / proxy interview
  • Register & survey/personal interview
  • Register & survey/proxy interview

• EU-SILC 2004-2007 (cross-sections)
Research Example – Lohmann (2011) - Findings

Table 1  Share of non-working population (%) with employee cash or near cash income

<table>
<thead>
<tr>
<th></th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>25–54 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>survey</td>
<td>4.1</td>
<td>5.1</td>
<td>4.8</td>
<td>6.2</td>
</tr>
<tr>
<td>survey-proxy</td>
<td>5.2</td>
<td>23.1</td>
<td>15.1</td>
<td>14.6</td>
</tr>
<tr>
<td>register</td>
<td>39.3</td>
<td>27.7</td>
<td>29.1</td>
<td>26.5</td>
</tr>
<tr>
<td>register-proxy</td>
<td>35.8</td>
<td>28.4</td>
<td>28.0</td>
<td>33.6</td>
</tr>
</tbody>
</table>


Note: Working defined as: person has worked at least one month during the income reference period

Source: Lohmann (2011: 42)
Research Example – Lohmann (2011) - Findings

• Are the group specific (working & non-working population) poverty rates affected by the type of data collection?

• Calculation of poverty rates for both criteria of working/ non-working population (employment or income based) by data collection approach
Figure 2  Absolute difference between at-risk-of-poverty rates of non-working/working population by country, data collection approach, year, and different criteria of ‘working’.

Source: Lohmann (2011: 46)
Research Example – Lohmann (2011) - Findings

• Further (multivariate) analysis

• Even though there are remarkable inconsistencies in the distributions of poverty measures depending on the data collection approach the general structure of poverty risk is not changed by the different approaches of data collection
Flexibility in terms of output harmonization

• Variables which should be delivered by the NSIs to Eurostat are precisely defined in regulations & guidelines

• Thus: variables included in the User Database are standardized according to the given rules

• However: National questionnaires are not standardized => while the output information is harmonised the input information must not be harmonised

  ➢ cross-national variations in the wording of questions or in the questionnaire design are very likely

  ➢ Risks of undocumented non-comparability in the data
Research example – Gash (2011)


• Research interest: How does unemployment affect social engagement? (together with Martina Dieckhoff)

• EU-SILC Module (2006) Social Participation

• Side aspect (data preparation): Compared the wording in the questionnaires: UK, FR, IRL, DE, DK, AU, SE
Research example – Gash (2011) – Main Findings

• Broad agreement in the questionnaire wording across countries, however:

• Some countries provide examples of the different types of social participation, while others provide none
  ➢ risk that countries without prompts will have lower observed social participation.

• Differences between countries in the social groups included in measures of social participation.
  • Some national surveys prompt that respondents should exclude *people they live with* while others don’t
  ➢ Risk of an upward bias in the latter case.
Source: Gash, Vanessa (2011): METHODOLOGICAL ISSUES in COMPARATIVE RESEARCH

- Other imperfections are such that the data is not comparable for some countries. i.e. Q: **PS090- Ability to ask relatives/friends/neighbours for help (y/n)**

<table>
<thead>
<tr>
<th>UK: I am going to describe two situations where people might need help. For each one, could you tell me if you would ask any of your neighbours for help?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) You are ill in bed and need help at home. Would you ask any of your neighbours for help?</td>
</tr>
<tr>
<td>(2) You are in financial difficulty and need to borrow some money to see you through the next few days. Would you ask any of your neighbours for help?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IRL/FR~/DE/DK: If the need arose would you feel able to ask a relative, friend or neighbour for help?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewer: Only relatives and friends (or neighbours) who don’t live in the same household as the respondent should be considered</td>
</tr>
<tr>
<td>1. Yes</td>
</tr>
<tr>
<td>2. No</td>
</tr>
<tr>
<td>3. Has no friends relatives or neighbours</td>
</tr>
</tbody>
</table>
Figure 6: Social isolation across EU countries: unable to ask any relatives, friend or neighbour for help (%), 2006

Source: EU-SILC Users’ database
NB: EU average: refers to the total population, 23 countries
Survey question: ‘ability to ask any relative, friend or neighbour for help’. The question is about ability for the respondent to ask for help irrespective of whether the respondent has needed it or not. Only relatives and friends who don’t live in the same household as the respondent are considered.
UK omitted (see Section 2)
Different types of reference periods

• Reference period: period of time to which a particular item of information relates.

• Two main types of reference periods in EU-SILC
  
  • income reference period
    ➢ all income information

  • current reference period
    ➢ most of other information

• further reference periods used in EU-SILC e.g.: Last twelve months; Since last year; since last interview; Childcare reference period
Different types of reference periods

• Income reference period
  • as a rule: a fixed 12-month period (such as the previous calendar or tax year)
  • exceptions:
    ➢ UK: current income (refer to the year of survey)
    ➢ IE: last 12 months

• Issues of concern
  • timeliness of income data
  • comparability across countries (UK, IE, all others)
  • differences in the reference period of income data & other data
    (current reference period = day of interview)
Different types of reference periods

• some variable available for both reference periods, e.g.

RX010: Age at the time of interview
RX020: Age at the end of income reference period

RB170: Main activity status during the income reference period
PL030: Self-defined current economic status

• But not all variables one might need when using income information are available or provided for both reference periods in the cross-sectional data

➢ People present at this year’s interview are not necessarily the same people present throughout the income reference period
### Issues of concern – Reference period

#### Example: Changes in household size, longitudinal data

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Decrease</th>
<th>No Change</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>2008</td>
<td>3.0%</td>
<td>94.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>2008</td>
<td>2009</td>
<td>3.7%</td>
<td>93.5%</td>
<td>2.8%</td>
</tr>
<tr>
<td>2009</td>
<td>2010</td>
<td>5.2%</td>
<td>90.5%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

Source: EU-SILC UDB_I10H_ver 2010-1 from 01-08-2012 (unweighted); own computation

* Household size is e.g. needed for the calculation of the equivalised household income
Need for more detailed information

- Identifiers to link cross-sectional & longitudinal data
  - Many variables are included in both files but not all
- More detailed information regarding substantial information
  - Isced
    - 6-level isced neglects the general/academic versus vocational divide
    - Substantial loss in the explanatory power of education
  - Income, e.g.
    - Some income variables are reported only at the household level but not at the individual level
    - Which household members receives them & how much each receives
  - Household, e.g
    - relationship HH-reference person to all other HH-members (included)
    - Partners and parents (included)
    - No household grid
Need for more detailed information

• More detailed information regarding methodological information, e.g.
  • All sample design variables should be included in the User Database

• Why?
  • EU-SILC is based on national samples (again: high diversity in the sample design between countries)
  • Population estimations are used to inform policy makers on poverty and social exclusion in Europe
  • Thus confidence intervals & standard errors are needed to indicate the precision of estimations
  • Which differences between countries are worthwhile further investigation which not?
Research Example – Goedemé, Tim (2010, 2013)


Issue of concern:

• Indicators which are measured at the household level (e.g. poverty, deprivation, work intensity) are often analyzed at the individual level (e.g. at-risk-of-poverty-rate of children)

➢ Standard errors might be underestimated if the clustering within households are not taken into account
Scenario 1: Assuming a simple random sample of individuals
Scenario 2: Assuming a sample of households, taking clustering within households into account

Source: Goedemé, Tim (2011): How much confidence can we have in EU-SILC? 2nd EU-LFS / EU-SILC User Conference. Mannheim, April 1, 2011
Finally ....

- No doubt there are more issues of concern

- The quality of EU-SILC data depends on the willingness and resources of the NSIs of continuously improving the data collection process as well as the data documentation

- Feedback of researchers to the NSIs & Eurostat might help

- Even though the focus was on issues of concern there is a large field of studies (not mentioned here) giving evidence of the benefits of EU-SILC

- As mentioned before: There is no perfect data in this world and the EU-SILC data is a unique data resource
  - It is in the responsibility of the researcher to ensure that the analyses he/she does is valid

- Thus the first (rather time-consuming but important) step in cross-national analyses are the reading of the data documentation as well the checking of the data
Thank you for your attention 😊

heike.wirth@gesis.org
### Average age of respondents by type of interview

<table>
<thead>
<tr>
<th>Year of the Survey</th>
<th>face to face interview-PAPI</th>
<th>face to face interview-CAPI</th>
<th>CATI, telephone interview</th>
<th>self-administered by respondent</th>
<th>proxy interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>49</td>
<td>50</td>
<td>47</td>
<td>44</td>
<td>41</td>
</tr>
<tr>
<td>2008</td>
<td>50</td>
<td>51</td>
<td>48</td>
<td>48</td>
<td>41</td>
</tr>
<tr>
<td>2009</td>
<td>50</td>
<td>51</td>
<td>48</td>
<td>46</td>
<td>42</td>
</tr>
<tr>
<td>2010</td>
<td>51</td>
<td>52</td>
<td>49</td>
<td>48</td>
<td>42</td>
</tr>
<tr>
<td>total</td>
<td>50</td>
<td>51</td>
<td>48</td>
<td>47</td>
<td>42</td>
</tr>
</tbody>
</table>

Source: UDB_l10P_ver 2010-1 from 01-08-2012 (unweighted), own computation
### Type of interview by sex

Example: Type of interview by sex of respondent

<table>
<thead>
<tr>
<th></th>
<th>face to face</th>
<th>CATI, telephone interview</th>
<th>self-administered by respondent</th>
<th>proxy interview</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PAPI</td>
<td>CAPI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>33,2%</td>
<td>24,0%</td>
<td>18,4%</td>
<td>24,4%</td>
</tr>
<tr>
<td>female</td>
<td>37,8%</td>
<td>28,5%</td>
<td>18,4%</td>
<td>15,2%</td>
</tr>
<tr>
<td>total</td>
<td>35,6%</td>
<td>26,4%</td>
<td>18,4%</td>
<td>19,6%</td>
</tr>
</tbody>
</table>

Source: UDB_l10P_ver 2010-1 from 01-08-2012 (unweighted), own computation