Introduction to the German Socio-Economic Panel (SOEP)

Joachim R. Frick
(DIW Berlin) <jfrick@diw.de>

March 2010
http://www.diw.de/gsoep

Special thanks to John P. Haisken-DeNew (RWI-Essen) for long-term cooperation in developing an earlier version of this introduction.
1. Introduction
2. Content
3. Sub-samples and Survey Issues
4. Sample Size Development
5. Recent Developments !!!
6. Data Structure
7. Preparing Data for Analysis
8. Dealing with Variables
9. Weighting
10. SOEP-Online Support
11. Data Distribution
12. Conclusion & Prospects

Appendix: NEWSPELL
“Division of Labor”: SOEP and official statistics

A rather European/German perspective

- **Short**-running panels by Official Statistics
  - ECHP, EU-SILC
- **Long**-running panel surveys under academic direction
  - SOEP, BHPS/Understanding Society, HILDA, PSID

Advantages

- Complementary contents, e.g.
  - In-depth income data in official panel surveys
  - Personal traits, health, subj. indicators in academic panels
- Mutual cross-validation
Well-known advantages of panel data

⇒ Decomposition of gross and net changes
   (e.g. poverty rate as a result of inflow, outflow and the cont. poor)

⇒ Causal Relationship / Evaluation
   (sequential ordering of “trigger events” and outcome)

⇒ Control for otherwise unobserved heterogeneity
   (RE and FE-models)
Increasing Opportunities for long-running panels

⇒ Cumulation of rare events by pooling (mortality, divorce, ...)

⇒ Increasing coverage of changes in institutional settings and the potential impact on individual behavior

⇒ Observation unit “household” guarantees appropriate coverage of births and deaths from cradle to crave

⇒ Linking objective outcome measures to subjective indicators (satisfaction as a proxy for utility)

⇒ Comparison of “intentions” and actual behavior (how relevant are expectations at the individual level?)

⇒ Intergenerational analysis linking parents and children

⇒ Increasing potential for cohort analysis

⇒ Improved efficiency in statistical modeling due to increased number of observations per individual
Potential Caveats (increasing with panel duration)

⇒ Bias due to selective attrition
⇒ Panel effects (incl. *positive* learning effects !!!)
⇒ Representative coverage of changes in underlying population due to immigration
⇒ “Compensation rules” for long-term shrinking number of observations
⇒ Continuity of questioning (phrasing of questions, modes, …)
  - inter-temporal comparability of indicators vs.
  - need for adjustments to cover institutional changes
SOEP Basics

What is “SOEP”?

- The SOEP was started in 1984: Now 24 Waves available!
- Longest-running longitudinal survey of private households and persons in the Federal Republic of Germany („Living in Germany“)
- Started with 6,000 Households in 1984, in 2007 approx. 12,000
- Over-sampling of Foreigners, Migrants
- East Germans, Top-up samples, Innovations, High Income

What are we measuring?

- Representative micro-data on persons, households, families
- Objective (e.g. income) and Subjective (e.g. satisfaction) indicators
- Measure Stability and Change in Living Conditions
- Topics in Economics, Sociology, Political Science, Psychology, Geography
- Retrospective Information on Biographical History
2. Content

Core Questions

- Population and demography
- Education, training, and qualification
- Labor market and occupational dynamics
- Earnings, income and social security
- Housing
- Health
- Household production
- Basic orientation (preferences, values, etc.)
- Satisfaction with life in general and various aspects
### Topic Modules ➔ Replication every ~ 5-10 years

<table>
<thead>
<tr>
<th>Year</th>
<th>Module Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>Employment biography</td>
</tr>
<tr>
<td>1985</td>
<td>Marriage and family biography</td>
</tr>
<tr>
<td>1986</td>
<td>Social origins (Bio), neighborh.</td>
</tr>
<tr>
<td>1987</td>
<td>Social security, early retirement</td>
</tr>
<tr>
<td>1988</td>
<td>Assets</td>
</tr>
<tr>
<td>1989</td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>1990</td>
<td>Time use and preferences</td>
</tr>
<tr>
<td>1991</td>
<td>Family and social services</td>
</tr>
<tr>
<td>1992</td>
<td>Social security and poverty</td>
</tr>
<tr>
<td>1993</td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>1994</td>
<td>Neighborhood, values</td>
</tr>
<tr>
<td>1995</td>
<td>Time use and preferences</td>
</tr>
<tr>
<td>1996</td>
<td>Family and social network</td>
</tr>
<tr>
<td>1997</td>
<td>Social security and poverty</td>
</tr>
<tr>
<td>1998</td>
<td>Ecology and envir. behavior</td>
</tr>
<tr>
<td>1999</td>
<td>Expectations, Use of time</td>
</tr>
<tr>
<td>2000</td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>2001</td>
<td>Social networks, working cond.</td>
</tr>
<tr>
<td>2002</td>
<td>Social Security, assets</td>
</tr>
<tr>
<td>2003</td>
<td>Ecology and envir. behavior</td>
</tr>
<tr>
<td>2004</td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>2005</td>
<td>Time use and preferences</td>
</tr>
<tr>
<td>2006</td>
<td>Family and social networks</td>
</tr>
<tr>
<td>2007</td>
<td>Social Security, assets</td>
</tr>
<tr>
<td>2008</td>
<td><strong>Further education or training</strong></td>
</tr>
</tbody>
</table>
Dimensions of Time

- Questions about a point of time (present)
  *e.g. current employment status or current levels of satisfaction*

- Single retrospective questions on certain events in the past (past)
  *e.g. how often did you change your job during the last ten years?*

- Retrospective life event history since the age of 15 (past)
  *e.g. employment or marital history*

- Monthly calendar on income and labor market issues (past)
  *e.g. employment status January through December last year*

- Questions concerning a period of time (past)
  *e.g. demographic changes since the last interview e.g. marriage*

- Questions concerning future prospects (future)
  *e.g. satisfaction with life five years from now, or job expectations*
Subsamples: Multi-step random sampling process

- **A** "West-German" residents: started in 1984, n=4,528 households
  Head is either German or other nationality than those in Sample B.

- **B** "Foreigners": started in 1984, n=1,393 households (oversampling)
  Head is either Turkish, Italian, Spanish, Greek, or Yugoslavian.

- **C** "East-Germans": started in 1990, n=2,179 households
  Head was a citizen of the GDR. (expansion of survey territory)

- **D** "Immigrants": started in 1994/95, n=522 households
  At least one HH member has moved to Germany after 1984.
  (expansion of survey population)

- **E** "Refreshment sample": started in 1998, n=1,067 households
  Random sample covering all existing subsamples (total population)

- **F** "Innovation sample": started in 2000, n=6,052 households
  Random sample covering all existing subsamples (total population)

- **G** "High Income sample": started in 2002, n=1,224 households
  Monthly net Household income > 7.500 DM (4.500 EUR in wave 2)

- **H** "Refreshment sample": started in 2006, n=1,506 households
  Random sample covering all existing subsamples (total population)
The SOEP sample 1984 – 2010 (households)
Mixed „Survey Mode“ in SOEP

Standardized, pre-tested instruments

- Face-to-face individual interviews with all HH-members 17+
  - Household interview with "head of household"
  - Paper-and-pencil interviews (samples A through D, E1, F, G)
  - No proxy interviews / (almost) no phone interviews
  - Self-Completers: Data Agency resolves inconsistencies

- Since 1998 stepwise implementation of CAPI
  - experimental design with split-samples E1/E2

- Prospective WAPI: self-admin. interviews via Internet (successfully tested)

- Important: support analyses of mode effects by adding information on mode to the micro-data!

Example: Interview Mode 1999/2002 (variable $PINTA$)

<table>
<thead>
<tr>
<th>Interview Mode</th>
<th>1999 (%)</th>
<th>2002 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Interview/Interviewer</td>
<td>43%</td>
<td>29%</td>
</tr>
<tr>
<td>Self-Completed w/o. Interviewer</td>
<td>29%</td>
<td>25%</td>
</tr>
<tr>
<td>Written (Snail-Mail)</td>
<td>14%</td>
<td>10%</td>
</tr>
<tr>
<td>Self-Completed w. Interviewer</td>
<td>5%</td>
<td>3%</td>
</tr>
<tr>
<td>CAPI</td>
<td>5%</td>
<td>28%</td>
</tr>
<tr>
<td>Part Oral / Part Self-Completed</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>Phone</td>
<td>.07%</td>
<td>.00%</td>
</tr>
<tr>
<td>Proxy</td>
<td>.04%</td>
<td>.04%</td>
</tr>
</tbody>
</table>
Development of the Interview-Mode

- PAPI
- CAPI
- SAQ (with Interviewer)
- SAQ (without Interviewer)
- SAQ (mail)
Survey Instruments: Address Log ($BRUTTO-files)

Containing general information (filled in by interviewer)

- on households (e.g. size, housing area, regional information);
- on individuals (e.g. sex, year of birth, relation to head)

- on the process of field work (e.g. number of contacts, reason for drop-outs, interview mode)

- different questionnaire versions according to survey status old ("green") v.s. new households ("blue")
**Survey Instruments: Questionnaires ($P/$H-files)**

- Pre-tested questions, only

- Yearly Standard Instruments
  - Household questionnaire
  - Individual questionnaire for each HH member aged 16 and over
  - Individual questionnaire „Gap“ (Temporary Drop-outs)
  - Until 1995 specific questionnaires for sub-samples
    (“Germans”, “Foreigners”, “East Germans”, “Immigrants”) as well as for survey status (“old” vs. “first time respondents”)

- Biography Questionnaire „Life History“
  - Until 2000 to be answered by first time respondents aged 17+
  - Covering life-time information up until 1st/2nd interview (education and labor market, marriage and fertility, socialization and parental background, immigration)
  - Since 2001 to be answered by first time respondents aged 18 and over
## 4. Sample Size Development

### Starting Sample Size in Wave 1 (full 100% sample)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Starting Year</th>
<th>Households</th>
<th>Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>A and B</td>
<td>1984</td>
<td>5,921</td>
<td>12,245</td>
</tr>
<tr>
<td>C</td>
<td>1990</td>
<td>2,179</td>
<td>4,453</td>
</tr>
<tr>
<td>D1/D2</td>
<td>1995</td>
<td>522</td>
<td>1,078</td>
</tr>
<tr>
<td>E</td>
<td>1998</td>
<td>1,067</td>
<td>1,923</td>
</tr>
<tr>
<td>F</td>
<td>2000</td>
<td>6,052</td>
<td>10,890</td>
</tr>
<tr>
<td>G</td>
<td>2002</td>
<td>1,224</td>
<td>2,671</td>
</tr>
<tr>
<td>H</td>
<td>2006</td>
<td>1,506</td>
<td>2,616</td>
</tr>
</tbody>
</table>
The Follow-Up Concept

⇒ All persons in HH are to be surveyed also the following years. At same address as well as after a residential move within Germany (⇒ regional mobility)
⇒ Personal interviews at age of 16 (⇒ demographic development)
⇒ Persons moving into an existing SOEP household. Since 1989 these persons are also followed in case of leaving the household. This had not been the case up to wave 5 (1988).

Temporary Drop-Outs

⇒ Principle: follow until two consecutive temporary drop-outs of all household members or a final refusal.
⇒ Gaps: small questionnaire including questions on central information which is missing for the year of the drop-out (e.g. employment status).
Demographic factors

- Persons exit by:
  - Death
  - Moving abroad

- Persons enter by:
  - Birth
  - Moving into a SOEP household from somewhere else in Germany or from abroad
  - Reaching age of 17 years (minimum respondents age is given by the calendar year, in which a person turns 17 years of age)
  - Split-offs of at least one old person from an old household
Field-work related factors (2 stages)

⇒ making a successful contact to a given household
⇒ realizing a successful interview

⇒ social groups typically associated with problems in respect to re-contacting and re-interviewing:
  • single person households
  • mobile households and persons
  • young adults leaving parental home
Cross-sectional perspective
⇒ Number of successfully interviewed households by sample
Cross-sectional perspective
⇒ Number of successfully interviewed persons by sample

[Bar chart showing the number of successfully interviewed persons for different samples from 1984 to 2008.]
Longitudinal perspective: Wave 1 response rates

Source: SOEP.
Longitudinal perspective: Wave-to-wave response rates
Longitudinal perspective: Sample A since 1984

Whereabout of the 11422 Persons

- Moved abroad
- Deceased
- Under the age of 16
- With interview
- Temporary drop-out
- Declined to reply
- No contact

Records without survey related attrition

Records with survey related attrition
Longitudinal perspective: Sample B since 1984

Whereabout of the 4830 Persons

- Moved abroad
- Deceased
- Under the age of 16
- With interview
- Temporary drop-out
- Declined to reply
- No contact
- Records without survey related attrition
- Records with survey related attrition
# The emergence of new households

<table>
<thead>
<tr>
<th>Persons</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
</table>
| Old     | • “classic case” without change of address  
          • entire household moves             | • Move-out                               |
| New     | • Birth                                  | • Birth                                  |
|         | • Move-in                                | • Caused by splitt-offs of old persons from old households* |

* Remember that households *new* to the SOEP may already have existed before contacting the survey
Incidence of New households after Wave 1

Household records in the SOEP: Wave 1 and Split-offs

As of 2008: Wave 25 (A and B); Wave 19 (C); Wave 15 (D1/D2); Wave 11 (E); Wave 9 (F); Wave 7 (G); Wave 3 (H).
What can we do to maintain high response rates in SOEP → „Governance“ of survey instruments and methodology

- Tailoring questions and survey instruments
  - What information needs to be collected annually?
  - Bi-annually collected information (Just asking for changes or do we need to ask explicitly for the current status? E.g. education)
  - Up to 5+ year intervals for very specific, complex issues (e.g. wealth)
- Incorporation of new areas of interest (e.g. psychological issues such as personality traits), allowing to collect „improved“ measures of the determinants of socio-economic behaviour
- Maintaining a fair balance between the necessary continuity of longitudinal aspects ("just stupid replication") and innovative developments in the social sciences and in survey methodology
- → Improving multidisciplinarity!
- → keeping respondents interested and involved
- → keeping users (researchers) actively involved
„Panel care“ – keep respondents involved

- Announcement by mail 2 weeks prior to interview
- For each successful interview, any respondent
  - receives a gift related to the yearly topical module (until 2007)
  - ticket for monthly nationwide lottery
  - since 2008: 5 € ex-ante incentive (until 2007: 1,50 € ex-post)
- Small presents for:
  - groups with higher response burden (mothers, first-time 17-yr-old respondents)
  - special gifts for kids in the household (balloons)
- Addresses are kept up to date by the field work agency during year
- Households receive the brochure „Living in Germany“ and information about data privacy regulations
- After interview (during summer): thank you-letter and „porto-card"
- Special treatments:
  - contact via phone
  - central case-by-case treatment for “problematic” households
  - mailing more information on request (e.g. Datenreport, a social report on Germany)
  - website “Leben in Deutschland” http://www.leben-in-deutschland.info/
- Highly qualified interviewer pool: Interview situation (F2F) ensures a personal relationship → harder to withdraw from the survey
  - Thus, the stability of the interviewer over time is very crucial
- Extensive phone follow-up by specially trained interviewers
- Offering multi-mode
„Multidisciplinarity“ – keep analysts involved

(1) Capturing the *full* life course
   - [before the] cradle → grave [and beyond]

(2) Observing the *unobservable*
The individual life-course in long-running household panel surveys

Household & family context

Birth

Individual life-course with prospective (yearly) data collection

Observations of subjective and objective variables (outcomes)

peer-networks, educational institutions, organizations, regional opportunity structure

social, cultural, and economic frame
Fostering longitudinal, intergenerational & life course research

• Improving data for **triggered** groups (and for cohort analyses) using **age- and event-specific instruments**
  
  ✓ Since 2003: data on new born children (about 0-15 mths)
  ✓ Since 2005: data on infants (about 3 yrs)
  ✓ Since 2008: data on first-graders (about 6 yrs)
  • Starting 2013: data on secondary level students (about 11 yrs)
  ✓ Since 2001: data on adolescence; first-time respondents (16/17yrs)
  ✓ Since 2009: Death of a spouse (Exit interviews)

• Event-triggered instruments (still in planning stage):
  • Divorce
  • Moving into retirement
Capturing the full life course

Improving measures about the early phase of life
→ „mother & child questionnaires“
The SOEP “Mother & Child Questionnaires”

Child related questions collected since 1984 (in household questionnaire) - Age specific questionnaires started in 2003 based on birth cohorts 2002/2003

<table>
<thead>
<tr>
<th>Age-specific questionnaires</th>
<th>Age-cohorts</th>
<th>Start (since)</th>
<th>Content</th>
<th>N (2008)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother-child-questionnaire I</td>
<td>ages 0-1</td>
<td>2003</td>
<td>Pregnancy, birth information, health of mother and child, temperament, care situation</td>
<td>1.436</td>
</tr>
<tr>
<td>Mother-child-questionnaire II</td>
<td>ages 2-3</td>
<td>2005</td>
<td>child health, temperament, activities with the child, care situation, adaptive behaviour (modified Vineland-Scale)</td>
<td>962</td>
</tr>
<tr>
<td>Mother-child-questionnaire III</td>
<td>ages 5-6</td>
<td>2008</td>
<td>child health, personality, activities of the child, care situation, socio-emotional behaviour (modified Strength and Difficulties Questionnaire)</td>
<td>237</td>
</tr>
</tbody>
</table>

Each wave about 200 additional children per age cohort....

- planning stage: age specific questionnaires for children in primary school (ages 8-9 in 2011), entering secondary school (age 10-12), etc. up to age 17 (adult questionnaire)
The SOEP “Mother & Child Questionnaire (I)”

4-page questionnaire for mothers who gave birth within the last 15 months

- Indicators about birth
  (Place, body measures: weight, height, head circumference)
- Health status of child
- Experience of pregnancy
  (physical and mental)
- Attitude towards child
- Partner’s support in daily child-rearing activities
The SOEP “Mother & Child Questionnaire (II)"

Questionnaire targeted at children aged 2-3 years

- Health status of child
- Temperament of child
- Activities of mother and child
- Child care situation
- Adaptive behavior (modified Vineland score (Sparrow et al 1984) with 4 domains (0-10 pts each):
  - verbal skills
  - activities of daily living
  - motor skills
  - social skills
  → summary score (0-40 pts)
The SOEP “Mother & Child Questionnaire (III)”

Questionnaire targeted at children aged 5-6 years

- Health status of child
- Personality of child
- Activities of mother and child
- Child care situation
- Use of media and language
- Socio-emotional behavior (modified Strength-and-Difficulties Questionnaire (SDQ), Goodman 1997)
  - Total difficulties score (0-30 pts)
  - Pro-social behavior (0-10 pts)

Application:
Berger & Spiess (2009): life satisfaction of mothers is positively correlated to children's skills and socio-emotional behavior
Capturing the full life course

Improving measures about the early phase of adulthood
→ „youth questionnaires“
The SOEP “Youth Questionnaire”

- Replaces standard biography questionnaire for „adults“ since 2000/01
- asked from all first-time respondents (17 yrs of age)
- Focus on age-specific issues:
  - Relationship to parents
  - Recreational activities (sports, etc)
  - School achievement
  - Planned educational attainment
  - Occupational prospects
  - Personality
  - Indicators for intergenerational mobility
The SOEP “Youth Questionnaire” (2)
[n=2,961 up to wave X, 2007]

Table 1: Target Population for the Youth Questionnaire by year, sample and age

<table>
<thead>
<tr>
<th>survey year / sample</th>
<th>A-E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>frequency</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>17 years</td>
<td></td>
<td></td>
<td></td>
<td>232</td>
<td>7.84</td>
</tr>
<tr>
<td>2001</td>
<td>17 years</td>
<td>17-19 years</td>
<td></td>
<td></td>
<td>618</td>
<td>20.87</td>
</tr>
<tr>
<td>2002</td>
<td>17 years</td>
<td>17 years</td>
<td></td>
<td></td>
<td>352</td>
<td>11.89</td>
</tr>
<tr>
<td>2003</td>
<td>17 years</td>
<td>17 years</td>
<td>17 years</td>
<td></td>
<td>365</td>
<td>12.33</td>
</tr>
<tr>
<td>2004</td>
<td>17 years</td>
<td>17 years</td>
<td>17 years</td>
<td></td>
<td>373</td>
<td>12.60</td>
</tr>
<tr>
<td>2005</td>
<td>17 years</td>
<td>17 years</td>
<td>17 years</td>
<td></td>
<td>368</td>
<td>12.43</td>
</tr>
<tr>
<td>2006</td>
<td>17 years</td>
<td>17 years</td>
<td>17 years</td>
<td></td>
<td>307</td>
<td>10.37</td>
</tr>
<tr>
<td>2007</td>
<td>17 years</td>
<td>17 years</td>
<td>17 years</td>
<td>17 years</td>
<td>346</td>
<td>11.69</td>
</tr>
</tbody>
</table>

Status: up to wave X (2007)
Capturing the full life course

Improving measures about the terminal phase of life
→ „exit interviews“
Realization of “Exit Interviews” in SOEP 2009 – triggering via personal questionnaire (1)

156. Has your family situation changed after December 31, 2007? Please indicate if any of the following apply to you and if so, when this change occurred.

<table>
<thead>
<tr>
<th>Event</th>
<th>Yes in 2008</th>
<th>Yes in 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>married</td>
<td></td>
<td></td>
</tr>
<tr>
<td>moved in with my partner</td>
<td></td>
<td></td>
</tr>
<tr>
<td>had a child</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Death in the household or in the family:
- My spouse/partner has died
- Father has died
- Mother has died
- Child has died
- Another person who lived here in the household has died

Please complete the questionnaire ‘The deceased person’

There have been no changes in my family
“Exit interviews” in SOEP (2)

Topics of the Questionnaire “The Deceased Person”
- Place and cause of death
- Health and illness in the last six months of life
- Relationship to the deceased person
- Satisfaction with life of the deceased person 3 months and 6 months prior to death

Limitations
- Almost no information about childless respondents (due to triggering)
- Statistical control of selectivity is necessary

Expected numbers of cases
- Information on about 150 deceased respondents out of about 200 who had died
- Multiple responses for about 30 deceased respondents

Survey instruments for capturing the life course

- Conception Phase
- Fetal Phase
- Childhood
- Teenage
- Adult Life

Birth
Terminal Phase
Death

- Individual Questionnaire Parents
- Questionnaire Mother & Child
- Household specific Questionnaire
- Teenager Questionnaire
- Individual Questionnaire

- Individual Quest. Partner and Children („Exit Intv.“)
- Address records

Memories & widow(er) Pensions
Challenge: Observing the „unobservable“
... and going beyond standard interview techniques

- Evaluation of health (by taking grip strength as a 1st step)
- Evaluation of cognitive abilities
- Personal traits (Big Five, risk aversion, trust, fairness, …)
Health (innovations since 2002)

✓ Standard questions since 1984 (self-rated health status, disability status, sick days, # nights in hospital, ...)

✓ Improved health measures since 2002

✓ SOEP-version of SF12 (physical and mental health status),
✓ Alcohol consumption (2006, 2008, 2010) (e.g. Ziebarth & Grabka 2009)
  ✓ Measurement by trained Interviewers (not too complex) using calibrated meters
  ✓ Two measures each hand (side information: left- vs. right-handed
  ✓ Grip strength = objective, reliable indicator of overall health
  ✓ Panel perspective: worsening of grip strength = indicator for reduced bodily capacity (labor market)
  ✓ Important complementary objective measure of health, alongside with the subjective health assessment
  ✓ → there is no significant indication for (selective) item-non-response and attrition (rather a welcome and interesting break in the standard interview situation)
Measuring Grip Strength in SOEP
Results of the SOEP-Pretest
Grip Strength by Age and Sex

First Measurement right hand

First Measurement left hand

Sex
- Men
- Women
Cognitive Abilities (since 2006)

- 2 Ultra short tests (90 sec) for adult respondents
- Full ability measures (about 30 min) for teenagers, only

Paper:

Anger & Heineck (2009): Do Smart Parents Raise Smart Children?: The Intergenerational Transmission of Cognitive Abilities, SOEPpapers #156
# Measuring Adults’ Cognitive Abilities

**Measures of cognitive ability:**
2 ultrashort tests (Lang, 2005)

<table>
<thead>
<tr>
<th>Symbol Correspondence test (fluid mechanics)</th>
<th>Word fluency test (crystallized pragmatics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- developed after the Symbol-digit-modalities-test (SDMT) by Smith (1973, 1995)</td>
<td>- developed after the Animal-naming-task (Lindenberger and Baltes, 1995)</td>
</tr>
<tr>
<td>- Respondents have 90 seconds to assign as many correct signs as possible to the consecutively displayed digits, while the appropriate assignment code is visible to them.</td>
<td>- Respondents have 90 seconds to name as many different animals as possible.</td>
</tr>
<tr>
<td>- This test measures fluid mechanics, which refers to general and largely innate ability.</td>
<td>- This test measures crystallized pragmatics, which includes all cognitive abilities that consist of accumulated knowledge and skills acquired in the past.</td>
</tr>
</tbody>
</table>
Measurement of Teenagers’ General Cognitive Abilities (DJ) (since 2006)

Objectives
• Gain insight in skill formation processes among young people (skill acquisition)
• Contribution to interdisciplinary research efforts on social disparities in educational attainment and labor market success

Intelligence-Structure-Test
• I-S-T 2000 R (Amthauer et al. 2001)

Dimensions
• Verbal potential (analogies)
• Numerical potentials (number sequences)
• Figural potentials (matrices/pictorial material)
• Reasoning (= sum score)

Survey
• Test take about 30 minutes
• Participation rate about 90%
• Additional incentive: stop-watch

Paper:
Personal Traits

- Big Five Inventory (2005, 2009): openness, conscientiousness, extraversion, agreeableness, neuroticism
- Reciprocity (2005, 2010): positive vs. negative
- Time preference (2006)
- Trust and fairness (2003, 2008)
- Affects (since 2007)
- Eagerness (2008)
- Impulsiveness (2008)

Selected applications:

# SOEP Survey Instruments in 2006

<table>
<thead>
<tr>
<th>Survey Instruments 2006</th>
<th>No. of cases</th>
<th>No. of pages (PAPI)</th>
<th>Duration (in minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address protocol</td>
<td>12,553</td>
<td>5</td>
<td>3-7</td>
</tr>
<tr>
<td>Household questionnaire</td>
<td>10,993</td>
<td>16</td>
<td>15-20</td>
</tr>
<tr>
<td>- Children (below age 16)</td>
<td>5,143</td>
<td>2</td>
<td>1-3</td>
</tr>
<tr>
<td>Individual questionnaire</td>
<td>20,049</td>
<td>36</td>
<td>35-40</td>
</tr>
<tr>
<td>Biographical questionnaire</td>
<td>223</td>
<td>20</td>
<td>20-25</td>
</tr>
<tr>
<td>Youth questionnaire</td>
<td>307</td>
<td>24</td>
<td>25-30</td>
</tr>
<tr>
<td>DJ - competency test</td>
<td>756</td>
<td>12</td>
<td>25-35</td>
</tr>
<tr>
<td>Mother &amp; Child I</td>
<td>234</td>
<td>4</td>
<td>5-10</td>
</tr>
<tr>
<td>Mother &amp; Child II</td>
<td>222</td>
<td>4</td>
<td>5-10</td>
</tr>
<tr>
<td>Grip strength (only 2006)</td>
<td>5,528</td>
<td>2</td>
<td>3-10</td>
</tr>
<tr>
<td>Cognitive short test (only 2006)</td>
<td>3,876</td>
<td>only CAPI</td>
<td>3-5</td>
</tr>
<tr>
<td>Temporary dropouts</td>
<td>Open</td>
<td>6</td>
<td>3-7</td>
</tr>
</tbody>
</table>
Cross-Sectional Data

Series of cross-sections within Panel Population

t₀  t₁  t₂

Drop-outs

Not yet in the sample or not yet interviewed
### Longitudinal Data ("wide format")

Complete case analysis with a balanced panel design

<table>
<thead>
<tr>
<th>t₀</th>
<th>t₁</th>
<th>t₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Longitudinal Data („wide format“)

Downstream model

<table>
<thead>
<tr>
<th>t₀</th>
<th>t₁</th>
<th>t₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Drop-outs
- Successfully interviewed in all waves
- New respondents
- Not yet in the sample or not yet interviewed
Longitudinal Data („wide format“)
Complete information analysis with an unbalanced panel design

- $t_0$
- $t_1$
- $t_2$

Drop-outs

Successfully interviewed in all waves

Not yet in the sample or not yet interviewed

New respondents
Longitudinal Data („long format“)

Pooling longitudinal data (two longitudinal datasets of two waves each)

- $t_0$
- $t_1$
- $t_2$

Drop-outs

Successfully interviewed in all waves

New respondents

Not yet in the sample or not yet interviewed
Principles of SOEP Data Structure

Defining the Population and Unit of Analysis

⇒ Survey Data are organized *cross-sectionally* (in principle)

⇒ Additional Longitudinal “meta files” are *explicitly* designed to support panel analyses
  - Defining population of interest: Gender, Age, Region, Samples
  - Defining time period: Years, Waves
  - Balanced vs. Unbalanced panel design
  - „Long“ (pooling) vs. „Wide“ format

⇒ Spell and Event Data
  - Data stored as Person-Events *not* Person-Years
$PBRUTTO$ 

Address Log

$P$ 

$PKAL$ 

$PAUSL^3$ 

$PLUECKE^2$

$POST^1$

$PAGE17^4$

$KIND$

Individual Level

Individual Questionnaire

Generated and Status Variables

$PEQUIV$

$PGEN$

$HBRUTTO$

Address Log

$H$ 

$EV$

GHOST

Household Level

Household Questionnaire

Generated and Status Variables

$HGEN$

$: Wave specification: A, B, C... Y

1 Waves G and H only; 2 Waves B through Q only; 3 Waves A through L only; 4 Starting Wave W.
## Data Structure: Longitudinal

### Individual/ Household
- **META-DATA**
  - PPFAD
  - HPFAD
  - PHRF
  - HHRF

- **WEIGHTING FACTORS**
  - PFLEGE

- **DROP-OUTS**
  - PBR_EXIT
  - PBR_HHCH

- **PERSONS NEEDING CARE**
  - PFLEGE

- **HEALTH**
  - HEALTH
  - GRIPSTR

- **WEALTH**
  - PWEALTH

- **SPELL**
  - OCCUP. BIO.
  - MARITAL STATUS

- **SOCIAL ASSIST.**
  - CALENDAR
  - (month)

- **INCOME**
  - ART.

- **HEALTH**
  - HEALTH

- **EIN.

- **PWEALTH**

- **HWEALTH**

### Cumulative Data
- **HOUSEHOLD**
  - OCCUP.
  - BIO.

### Individual
- **BIRTH**
  - (women & men)

- **PARENTAL INFO**
  - BIO.
  - BIRTH

- **AGE SPEC. INFO**
  - BIOSOC
  - BIO-AGE17

- **MULTIPLES**
  - BIOTWIN

- **MIGRATION**
  - (migrants only)

- **FIRST JOB**
  - BIO-

- **IMMIG**

- **JOB**
**Data Structure: Matching**

- **Meta Data**
- **Wave 1**
- **Wave 2**
- **Wave 3**

- "filled gap" of **Net** Population (longitudinal) = balanced panel design

- "Gross" Population (longitudinal) = balanced panel design

- "Net" Population (longitudinal) = unbalanced panel design

- **Universe of SOEP Population**
- **Cross-sectional Population**
- □ "Gross" Population
- □ "Net" Population
Data Structure: Matching

Household Files

- HPFAD multi-wave household meta information
- HHRF household weights
- $HBRUTTO single-wave household address register
- $H/$HOST single-wave household data
- $HGEN single-wave generated household data
- GGKBOU* regional information: community size classification

ID (Matching Vars): HHNR, HHNRAKT ($HHNR)

* not available in Scientific Use Version for users outside the EU (Data Privacy), however, access possible via SOEPremote)
## Data Structure: Matching

### Person Files

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPFAD</td>
<td>multi-wave person meta information</td>
</tr>
<tr>
<td>PHRF</td>
<td>person weights</td>
</tr>
<tr>
<td>ID (Matching Vars):</td>
<td>HHNR, PERSNR</td>
</tr>
<tr>
<td>$PBRUTTO</td>
<td>single-wave person register</td>
</tr>
<tr>
<td>$P/$POST</td>
<td>single-wave person data</td>
</tr>
<tr>
<td>$PGEN</td>
<td>single-wave generated person data</td>
</tr>
<tr>
<td>$PKAL/$PKALOST</td>
<td>single-wave person monthly calendar (JAN-DEC)</td>
</tr>
<tr>
<td>$PAUSL</td>
<td>single-wave foreigner data (Sample B, only)</td>
</tr>
<tr>
<td>$PLUECKE</td>
<td>single-wave person gap data</td>
</tr>
<tr>
<td>$KIND</td>
<td>single-wave child data</td>
</tr>
<tr>
<td>$PEQUIV</td>
<td>CNEF Variables + background variables</td>
</tr>
<tr>
<td>ID (Matching Vars):</td>
<td>HHNR, HHNRAKT ($HHNR), PERSNR</td>
</tr>
<tr>
<td>PBR_EXIT</td>
<td>multi-wave person register (demogr. drop-out)</td>
</tr>
<tr>
<td>PBR_CGHH</td>
<td>multi-wave person register (HH changes)</td>
</tr>
<tr>
<td>ID (Matching Vars):</td>
<td>HHNR, PERSNR, SVYYEAR</td>
</tr>
</tbody>
</table>
Data Structure: Matching

Biography & Spell Data

- BIOIMMIG: Migration Biography
- BIOJOB: First Job, Entrance to Labor Market
- BIOBIRTH/M: Multi-wave birth biography
- BIOPAREN: Parental information
- BIOAGE01: Data on "mother & child" (around birth)
- BIOAGE03: Data on "infants" (2-3 years old)
- BIOAGE06: Data on "first-schoolers" (5-6 years old)
- BIOAGE17: Youth data (16-17 years old)
- BIOSOC: Data on adolescence
- PBIOSPE: Empl. biography (yearly data since age 15)
- ARTKALEN: Activity calendar (monthly data, 1983-2007)
- BIOMARSM: Marital status by month
- BIOMARSY: Marital status by year
- SOZKALEN: HH-based social assistance calendar

ID (Matching Vars): HHNR, PERSNR, (SPELLNR/ SVYYEAR)
Cumulative Data

- HWEALTH  Wealth Information HH Level (2002, 2007)
- GRIPSTR  Physical Health Measure Grip-strength (2006)
- PFLEGE   Persons in need of care (invalidity) data (1985-2008)

ID (Matching Vars): HHNR, PERSNR/HHNRAKT, SVYYEAR

!!! Forthcoming: release of a SOEP database in „long-format“ (all files containing cross-sectional information pooled across all years using longitudinally consistent var names and values)
Event and Spell Data

Although in the course of time the absolute number of observations (e.g., individuals) is almost steadily decreasing from a cross-sectional perspective, the cumulative number of events and/or spells covered by the entire data is increasing wave by wave.

Events

- Re-migration: YPBRUTTO (up to 2007: >1,600 events)
- Deaths (mortality): YPBRUTTO (up to 2007: >2,800 events)
- Births (fertility): BIOBIRTH (up to 2007: 32,000 births with approx. 21,000 of these persons being identified within the SOEP population)

Spells

- Monthly and yearly labor market status: ARTKALEN, PBIOSPE (as of 2007: e.g. 20,000 spells of unemployment, >57,000 FT employment)
- Monthly and yearly marital status biography: BIOMARSM, BIOMARSY
- Monthly receipt of social assistance: SOZKALEN (household)
Status Variables

⇒ Problem:
  • Some information is collected in the first interview only.
  • Old respondents are asked for changes since last year’s interview, while new respondents have to fill in the current status.

⇒ Example: since when with current employer

⇒ Solution:
  • The collected information is stored in different variables.
  • In order to minimize computing efforts for the user, the GSOEP provides yearly **status variables** on individual and household level, which integrate all of these information in a common variable showing the current status for all respondents.
  • Thus, there is nothing else but a re-organisation of already existing data and there is almost no assumption or normative setting involved in the generating process.
  • Variable $ERWZEIT in file $PGEN gives the number of years with the current employer for all employed respondents

Data Files: $PGEN, $HGEN, PPFAD, BIO*
Generated Variables

- **Problem:**
  Provide a user-friendly single output variable

- **Example:** Typically required institutional fulltime years of education

- **Solution:**
  Variable $BILZEIT$ in file $PGEN$: Based on separate variables measuring schooling attainment, completed training certificates, vocational education, university degrees, etc., a total number of years of education is generated

Data Files: $PGEN$, $HGEN$, PPFAD, BIO*
# 8. Dealing with Variables

## Principles for Naming Survey Variables

<table>
<thead>
<tr>
<th>Digit</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wave</td>
</tr>
<tr>
<td>2</td>
<td>Unit of analysis: P=individual, H=household</td>
</tr>
<tr>
<td>3-4</td>
<td>Number of question in survey instrument (questionnaire)</td>
</tr>
<tr>
<td>5-6</td>
<td>Number of item in survey instrument (questionnaire)</td>
</tr>
<tr>
<td>5 or 7</td>
<td>Differentiation according to sub-sample: A=Foreigners, O=East Germans</td>
</tr>
<tr>
<td>2-8</td>
<td>Text for Variables in $BRUTTO files and some occupation-related variables in $P files</td>
</tr>
</tbody>
</table>

### Examples:

- **AP04** Wave 1; Individual; Question 4
- **BH0502** Wave 2; Household; Question 5; Item 2
- **DP24G09** Wave 4; Individual; Question 24; Green version; Item 9
- **AP64A** Wave 1; Individual; Question 64; Sample B, Foreigners
- **BIS88** Wave 2; Intl. Standard Classification of Occupation (ISCO88)

### Exceptions:
- Identifiers (PERSNR, HHNKR, HHNRAKT, SVYYEAR);
- generated vars (files $PGEN, $PKAL, $HGEN, BIO*, $PEQUIV)
Dealing with Missing Values

- SOEP data differentiates three kinds of missing values:

<table>
<thead>
<tr>
<th>Code</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>no answer / do not know / item non-response</td>
</tr>
<tr>
<td>-2</td>
<td>does not apply</td>
</tr>
<tr>
<td>-3</td>
<td>after checking for plausibility a given value was found to be implausible and finally deleted (to be interpreted like -1)</td>
</tr>
</tbody>
</table>

- However, there are NO system-missing values!
**Imputation**

**INR** (codes “–1” or “–3” in original variable) and **PUNR**

**Income:**
- Single imputation of missing income variables due to INR using...
  - „Row-and-Column“ imputation based on longitudinal data (Little & Su, 1989)
  - x-sectional imputation (mostly regression-based) if no panel data available
  - Annual income (wrt previous year; stored in file $PEQUIV)
  - Current monthly labor income (stored in file $PGEN: LABGRO$$, LABNET$$)
  - Imputation flags indicating imputation status
- Since release 2009: single imputation of PUNR (using longitudinal data)
- Since release 2008: multiple imputation for current monthly HH net income („screener“); stored in file $HGEN and MIHINC

**Wealth:**
- release 2007: multiply imputed data on individual wealth (2002) stored in files PWEALTH (incl. PUNR) and HWEALTH (household level wealth aggregates)
  - wealth data for 2007, 2012, etc. will be cumulatively stacked into these files
  - wealth data release for 2007: multiple imputation using *longitudinal data*

**Housing:**
- release 2009: imputed data on housing variables (all waves) stored in $HGEN
9. Weighting

Why is Weighting Necessary?

⇒ Different design probabilities for subsamples A-G in wave 1

⇒ Unit-Non-Response
  • not willing to participate in the first wave
    • 1984 subsamples A and B
    • 1990 subsample C
    • 1994 subsample D1 / 1995 subsample D2
    • 1998 subsample E
    • 2000 subsample F
    • 2002 subsample G
    • 2006 subsample H
  • attrition in the subsequent waves due to
    • unsuccessful follow-up (lost contact)
    • refusal (temporary or final)
Sampling Information Across Subsamples
Cross-Sectional Weighting Wave 1

<table>
<thead>
<tr>
<th>Subsample</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Foreigners in West-Germany&quot;</td>
<td>4.500</td>
<td>1.500</td>
<td>2.000</td>
<td>500</td>
<td>1.000</td>
<td>6.000</td>
</tr>
<tr>
<td>&quot;German Residents in the GDR&quot;</td>
<td>61</td>
<td>68</td>
<td>70</td>
<td>55-75</td>
<td>54</td>
<td>52</td>
</tr>
<tr>
<td>&quot;Immigrant since 1984&quot;</td>
<td>.0002</td>
<td>.0008</td>
<td>.0004</td>
<td>.0002</td>
<td>.0003</td>
<td>.00016</td>
</tr>
<tr>
<td>&quot;Refreshment&quot;</td>
<td>5,000</td>
<td>1,250</td>
<td>2,500</td>
<td>5,000</td>
<td>33,333</td>
<td>6,000</td>
</tr>
</tbody>
</table>

Weight = $1 / [ P(D_i=1) \times P(R_i=1|D_i) ]$
External Information: Adjust Wave 1 Weights

- **Household:** Sex of head, Age of head, Size of household, Nationality of head, Country of origin of immigrants

- **Resident population in private households:** Sex, Age, Marital, Status, Nationality of head

- **Resident population, total:** Sex, Age, Counties, Size of community

- **Children up to 16 years of age:** Sex, Type of School, Nationality of head

- **Gainfully employed persons residing in private households:** Sex, Age, Job (ISCO-1 digit), Nationality
Longitudinal Weights

⇒ **Idea:** Estimate Probability to remain in sample: \([ t, t+1 ]\) Separately for Each subsample (Logit Regression)

⇒ Contact Probability (CP)

⇒ Response Probability (RP), given Contact

⇒ Inverse Staying Probability = \(1 / [CP \times RP]\)

⇒ **Longitudinal Weight** = \([\text{Cross-Section Wgt Starting Wave}] \times [\text{Inv Staying Prob}]\)
Significant Predictors for Contact Probability

- Whether Moved
- Large City
- Household size
- Type of house
- Split-off household

Significant Predictors for Response Probability

- Age/Gender of household head, Type of the household
- Change of interviewer, Number of interviews
- Household head Member of Wave 1
- Person moving out, Marital status, Separation of a couple
- Unemployment, Expected Job Loss, Occupational status
- Welfare Recipient, Household income
- Household income not reported, Assets, No assets reported
- East-West Migration, Migrant, Telephone, Sub-tenant
Ready-Made Variables Provided

- **Cross-Section**: Person ($PHRF$) and Household ($HHRF$) level
- **Inverse Staying Probability**: Person and Household level
- **Sample Specific Cross-Section Weights** (Samples D, F, G)
- **Separate weights for time series** (e.g. on income inequality) dropping wave 1 of a given sample ($PHRF1$)
- **Support for Variance Estimation**: Random Groups, Strata

Data Files: PHRF, HHRF, DESIGN
10. Online Support

SOEP Homepage
- http://www.diw.de/gsoep
- Online Support Services, Documentation, FAQ, SOEP newsletter
- link to SOEP-Research Data Center (SOEP-RDC)

SOEP info
- Information System: Frequencies (unweighted), Questionnaires
- Getting Started Quickly: Generate Command Files!

Desktop Companion (DTC)
- In-depth Description in English
- Basic, Extensions, Retrievals, Weighting

SOEP lit
- Database of articles, books, papers written using SOEP
- Easy Search Interface

SOEP monitor
- Time series on wide range of cross-sectional and longitudinal indicators of living conditions in East and West Germany (weighted)
- Benchmark for users!!
http://www.diw.de/gsoep

Online Information

SOEP info
SOEP lit
DTC
Newsletter
Links...

http://www.diw.de/gsoep
WWW-SOEPinfo is an interactive system with extensive search capabilities that provides you with detailed information on the variables in the SOEP dataset. By interactively creating lists of variables in SOEPinfo, the user can then output frequencies information, item correspondence, and even generate SPSS, SAS, and Stata command files.

SOEPinfo should be your first stop in starting any new SOEP project!
Desktop Companion (DTC)

Main Resource

Written in LaTeX, automatic Adobe PDF Bookmarks

Clickable
SOEPlit - Literature

Search for all DP's, Articles, Books using SOEP

Input any Keyword, use Boolean Operators
11. Data Distribution

⇒ Strict data protection legislation requires ... 
  • Data user contract with DIW Berlin 
  • Intl. Users: Co-operation with Cornell Univ., Ithaca/NY

⇒ Data dissemination on CD-Rom / DVD only 
  • detailed written documentation 
  • full German and English labeling (variables and values) 
  • within EU: 100% sample 
  • outside EU: International Scientific Use Version (95% sub-sample) 
    • detailed regional info not available

⇒ Low nominal data fee (30 EUR/year; 125 USD one-time) 
  • yearly updates on CD-Rom / DVD

⇒ Access to sensitive information (e.g. regional data): 
  • SOEPreomote allows processing of sensitive data from outside DIW Berlin
Data Distribution – DVD

Windows Setup Program

Ready-to-Go

⇒ SPSS
⇒ Stata
⇒ SAS
⇒ ASCII

File Format
- STATA
- SAS XP'T
- SAS (binary)
- SPSS (SAV files)
- SPSS (PQR files)

Select Files
- Select All
- Unselect All

Select Waves A-R (1-18) or Other Files

Current Archive

STA_100.EXE

221 files selected

Define Directories
Quit

INSTALL THE FILES!

DIW BERLIN
Geo-codes & Remote Data Access

Geo-coded data at various levels
- county level (Landkreis)
- community level (Gemeinde)
- ZIP-code level (official geo-coded data)
- block level (commercial data)

Linking Neighborhood Information
- based on geographic coordinates (latitude and longitude)

Access via SOEPremote
Cross-National Equivalent File (CNEF)

- Standardization of Various Measures
  (Focus: Income, Demographics, Employment, Health)

- USA PSID - Panel Study of Income Dynamics
- Germany SOEP - German Socio-Economic Panel Study
- Great Britain BHPS - British Household Panel Study
- Canada SLID - Survey of Labour and Income Dynamics
- Australia HILDA - Household, Income and Labour Dynamics in Australia
- Switzerland SHP - Swiss Household Panel

12. Conclusion & Prospects

Due to recent and upcoming innovations …

- SOEP stands for theory based, research driven data collection – not just “more and better statistics”
- Improved potential for analyses of smaller subpopulations due to increased sample size (~ 12,000 HH since 2000)
- Improved potential for intergenerational analyses based on more than 20 waves of data (PSID ~ 35th wave!)
- Improved biographical data (e.g., youth and adolescence)
  - relevant background information (RHS-variables)
  - as well as of self-contained interest
- Improved data on methodological issues (interview modes, imputation of missing values (income & wealth), …)
- Improved geo-coded data (accessible via SOEPremote)
- Better controls for otherwise unobserved heterogeneity (behavior, health, personal traits, etc.)
SOEP’s international networking improves the cross-national infrastructure for data and analyses

- Cross-sectional databases: LIS, LWS
- Longitudinal databases:
- Development of user-friendly and cross-nationally comparative micro-data (ex-post harmonization)
  - education: ISCED, CASMIN ($PGEN$)
  - labor market: ISCO88, NACE ($PGEN$)
  - regional information: NUTS ($HGEN$)
  - income: Canberra Group recommendation ($PEQUIV$)
International collaboration …

- with other producers and analysts of panel data (e.g. BHPS, HILDA, PSID)
- and with respect to data collection (e.g. pre-testing),
- methodology (e.g. weighting, imputation),
- and substantive issues (e.g. timing of special topical modules)

→ simplifies future (ex-ante) data harmonization and quality of data for cross-national analyses


Thank you for your attention!

E-mail: jfrick@diw.de
SOEP-Hotline: soepmail@diw.de
SOEP-Homepage: www.diw.de/gsoep
Aim: Minimizing and Controlling for Sampling and Measurement Error in a Longitudinal Context

- **Sample Selection**
  - Compensation for Non-Response
    - Weighting, Multiple Imputation (Spieß & Goebel 2005; Frick & Grabka 2005)
  - Reduction of Non-Response
    - Incentives, Response Burden, Interview Settings (Schräpler 2004, 2006)

- **Measurement Instruments**
  - New Modes of Data Collection
    - Computer-Assisted Interviewing, Web-Surveys, Mobile Phones (Schräpler et al. 2006; Kroh 2005)
  - New Response Formats
    - Grip-Strength, Experiments (also: Factorial Designs) (Hank et al. 2008; Fehr et al. 2002, Poster Auspurg et al.)
  - Multiple-Item Scale Construction
    - Personality, Health, Cognitive Ability (Gerlitz & Schupp 2005; Lang et al. 2007)

- **Respondent and Interviewer Behavior**
  - Response-Artifacts
    - "Classical" Questionnaire Design, Panel Effects (Kroh 2007; Frick et al. 2006)
  - Interviewer-Effects
References: Sample Selection

**Compensation for Non-Response**

**Reduction of Non-Response**
References: Measurement Instruments

- **New Modes of Data Collection**

- **New Response Formats**

- **Multiple-Item Scale Construction**
References: Respondent and Interviewer Behavior

- **Response-Artifacts**

- **Interviewer-Effects**
Aims of NEWSPELL

- Creation of distinctive calendars
- Aggregation of events
- Disaggregation of combined events
- Re-Definition of time range (begin and end)
Appendix: NEWSPELL

Data supported by NEWSPELL

Existing SOEP-Data (use with current SOEP-password)

- PBIOSPE.DAT  Combined activity calendars by year
- ARTKALEN.DAT  Activity calendars by month
- EINKALEN.DAT  Income calendar by month
- BIOMARSY.DAT  Marital status by year
- BIOMARSM.DAT  Marital status by month
- SOZKALEN.DAT  Social assistance by month (households)

Self-defined Spell-Data (free use without password)
Appendix: NEWSPELL

Output of NEWSPELL

- LOG-File keeps a record of your session
- data in spell-format with additional variables
  - previous spell-type
  - next spell-type
  - censoring status
- data in time-series format
Appendix: NEWSPELL

How to use NEWSPELL?

- Copy all files from folder \NEWSPELL on your hard-disk
- Write a simple command-file with any DOS-editor
- Start the program with

  NEWSPELL command-file [password]
Appendix: NEWSPELL

Example:

original spell-data

case persnr spellnr spelltyp begin end
  1  1001  1  2  1   8
  1  1001  2  1  9  121  1001  3  3  8   9
  1  1001  3  3  8   9
  1  1001  4  3  12  14
  1  1001  5  1  16  18

/* type 1 = full employed
/* type 2 = part time employed
/* type 3 = unemployed
Appendix: NEWSPELL

command file

NI=example.dat /* a self defined dataset (input)
NS=example.spl /* output in spell-format
NT=example.tim /* output in time series format
NL=example.log /* log-file of session
NR=example.res /* output-file w/ frequencies of new vars

/* definition of new spell system
/* new type 1 = employed and unemployed
/* new type 2 = only employed
/* new type 3 = only unemployed

1=1 and 3 or 2 and 3 /* disagg. and aggregation: employed and unemployed */
2=1 or 2 /* aggregation
3=3 /* unemployed

NB=2 /* New range: Begin
NE=14 /* New range: End
## Appendix: NEWSPELL

### Output data files:

#### time-series data:

<table>
<thead>
<tr>
<th>case</th>
<th>persnr</th>
<th>begin</th>
<th>end</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
<th>#11</th>
<th>#12</th>
<th>#13</th>
<th>#14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1001</td>
<td>2</td>
<td>14</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

#### spell data:

<table>
<thead>
<tr>
<th>case</th>
<th>persnr</th>
<th>spellnr</th>
<th>type</th>
<th>begin</th>
<th>end</th>
<th>previous</th>
<th>next</th>
<th>censor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1001</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>-1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>1001</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1001</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1001</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1001</td>
<td>5</td>
<td>3</td>
<td>13</td>
<td>14</td>
<td>1</td>
<td>-1</td>
<td>2</td>
</tr>
</tbody>
</table>

#### censor: (1) uncensored (2) right-censored (4) left-censored (5) l+r-censored