Levels, distribution and drivers of lifetime earnings

Linkage project SOEP-RV

Mattis Beckmannshagen, Timm Bönke, Martina Kaplanová,

Carsten Schröder, Yogam Tchokni

DIW Berlin, FU Berlin

September 29, 2023

Outline

1. Context
2. Data
3. Method
4. Result
5. Conclusion

Motivation I

Most empirical analysis of inequality: snapshots of economic resources for a selected sample

Lifetime earnings inequality literature: focus on men with high labor force attachment [Guvenen et al., 2021, Bönke et al., 2015]

Limitations:

- heterogeneous earnings dynamics over the life course
- role of household redistribution in earnings inequality
- substantial portion of the work force not included in analysis

Motivation II: Data

Data used for inequality research:

1. Administrative data: great earning records, limited background information
2. Survey data: self-reported and limited earnings records, great background information

Combining the two data sources:

→ allows thorough analysis of drivers of lifetime earnings inequality

→ enables to study a larger population as missing labor market information can be explained

Research question

1. How did lifetime earnings develop across people born between 1935 and 1973?

2. How did inequality in lifetime earnings develop for men and women across the same birth cohorts?

3. What are the exogenous drivers determining one's position in the earnings distribution? *[Work in progress]*

Data I: SOEP

German Socio-Economic Panel

- Annual survey data of 15 thousand households and 30 thousand individuals in Germany starting 1984
- Individual and household level information
- Crucial for us:
  - employment history
  - partner and children information
  - parental background
  - geographic location
  - health indicators

Data II: DRV

German Pension Insurance (DRV)

1. *Fixed part*: Time invariant individual-level information
2. *Variable part*: Monthly points from social security

Data preparation steps:

- *Pareto imputations* of high earners: earnings right-censored
- *Imputation of civil servant and self-employed earnings*: not reported in DRV, imputed using SOEP reported earnings

Data III: SOEP-RV

- The SOEP collects comprehensive information about many life domains at individual and household level. Unfortunately, because of panel attrition, biographies are incomplete.
- The German Pension Insurance provides complete insurant biographies. Unfortunately, the information is individual level only and focusing on a single life domain, insurance.

The record-linked SOEP-RV dataset combines the strengths of both datasets.

Sample size and restrictions

Restricted to:

1. Birth year: 1935 to 1973
2. Geography: never worked in the former GDR
3. Labor market attachment: worked min 1 year in lifetime

	Individuals	Observations
Asked for consent	23,145	—
Consented	12,298	—
Records matched	12,054	—
In sample	9,865	219,074
Reach age 45	5,366	150,248

Table: Summary of SOEP-RV record linkage

Selectivity & other issues

Comparison our sample of linked cases from the SOEP-RV with the overall pension insurance records (RV-VSKT sample)

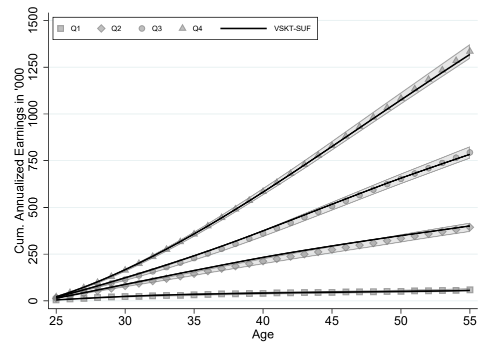


Figure: UAX mean cumulative earnings by quartile: SOEP-RV vs RV-VSKT

Unknown zeros

Missing vs zero earnings: without further information, administrative data cannot effectively distinguish the reasons for missing earnings

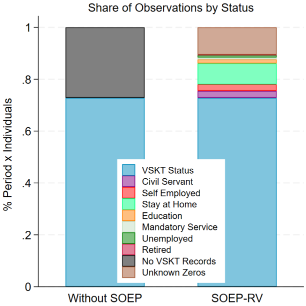


Figure: Unknown zeros: SOEP-RV vs RV-VSKT

Earnings concepts

Annualized lifetime earnings: earnings from age 25 to age 45:

$$\overline{Y}^i \equiv \frac{1}{21} \times \sum_{t=25}^{45} Y_t^i \tag{1}$$

Two earnings concepts:

1. **Standard earnings**: only from **dependent** employment, **min 10 years** in labor market
2. **Augmented earnings**: from all employment[1], **min 1 year** in labor market

[1] Including civil servants and self-employed

Inequality metrics

Theil index:

$$I = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{\mu} \ln \left(\frac{y_i}{\mu} \right) \tag{2}$$

Decomposition into between- and within-group inequality:

$$I = \sum_{k} = 1^m s_k I_k + \sum_{k} s_k \ln \left(\frac{\mu_k}{\mu} \right) \tag{3}$$

where

- $s_k = \frac{n_k \mu_k}{n\mu}$ is the share of earnings in group k, and
- μ_k is the average earnings in group k, and
- μ is the total average earnings

Drivers

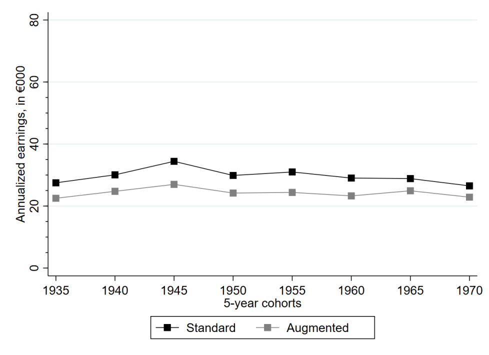
We use a battery of **exogenous characteristics** from the SOEP to assess the drivers which determine one's position in the earnings distribution

Analysis underway:

- Binary classification exercise: predicting top 10% lifetime income at age 45 in cohort, by gender
- Testing Logistic Regression Model and Random Forest Classifier
- Interpretation of RF Model using different measures:
  - Feature importance (Gini Impurity, Permutation, SHAP)
  - Partial Dependence

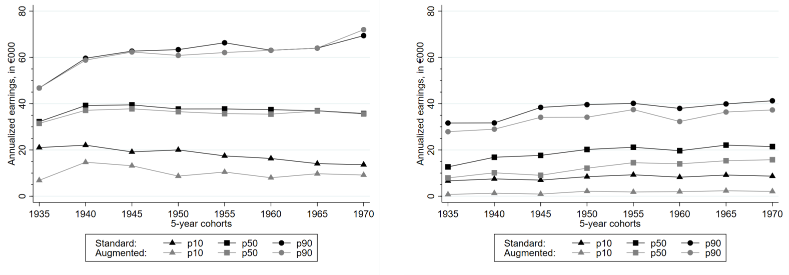
Earnings levels I: Median earnings for both genders

CPI-adjusted lifetime earnings: minimal changes over the eight 5-year cohorts, despite the economic boom in post-war West Germany



Earnings levels II: Median earnings by gender

- **Men**: flat median hides increasing earnings for top-earners and decreasing earnings for bottom-earners
- **Women**: increasing earnings for women along the earnings distribution, more so in the upper tail

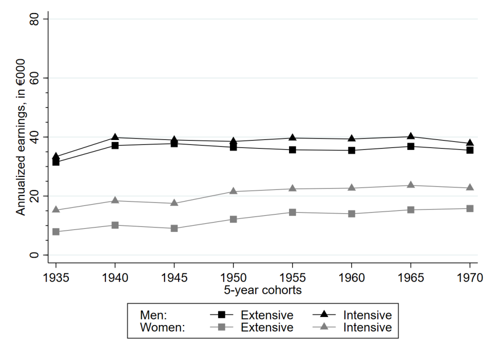


(a) Men

(b) Women

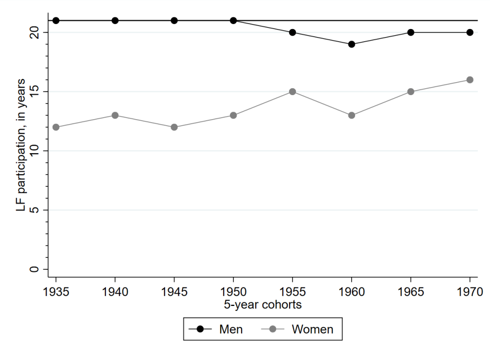
Earnings levels III: Extensive and intensive margin of earnings

- Extensive: cumulative earnings divided by 21
- Intensive: cumulative earnings divided by number of years in the LM



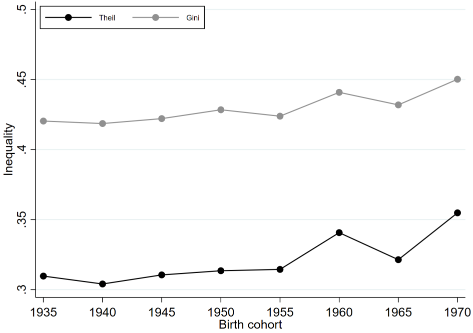
Earnings levels IV: LF participation by gender

- **Men**: decreasing labor force participation, still much higher than for women
- **Women**: increasing labor force participation for subsequent cohorts



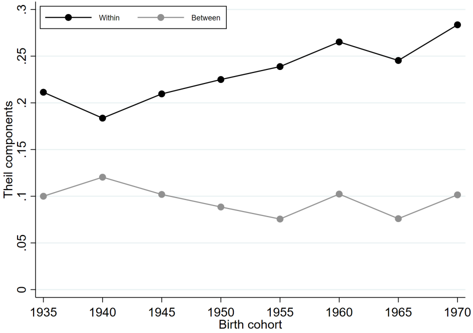
Inequality I: Gini vs. Theil Index

- Flat for older birth cohorts
- Increasing for later-born cohorts



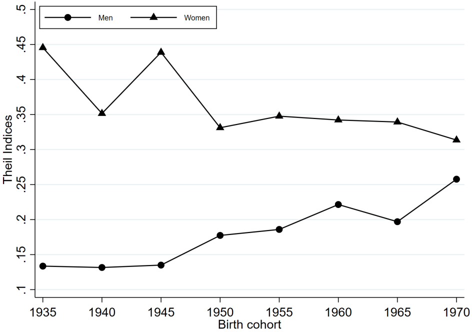
Inequality II: Theil between and within components

- Between inequality slightly decreasing
- Within inequality increasing across cohorts



Inequality III: Theil index within genders

- Women: slightly decreasing inequality
- Men: Sharply increasing in older cohorts



Drivers: Descriptives

Moving up the income distribution, the following characteristics are more/less prevalent:

Variable	**Men**	**Women**
Height	+	·
Migration background	−	−
Single parent	−	·
Siblings	−	−
Mother highly educated	+	·
Father highly educated	+	+
Mother in blue collar occupation	·	+
Mother self-employed	−	−
Mother civil servant	+	+
Father in blue collar occupation	·	·
Father self-employed	−	−
Father civil servant	+	+
Grew up in rural area	−	−

Drivers: Descriptives

Moving up the income distribution, the following characteristics are more/less prevalent:

Variable	**Men**	**Women**
Height	+	·
Migration background	−	−
Single parent	−	·
Siblings	−	−
Mother highly educated	+	·
Father highly educated	+	+
Mother in blue collar occupation	·	+
Mother self-employed	−	−
Mother civil servant	+	+
Father in blue collar occupation	·	·
Father self-employed	−	−
Father civil servant	+	+
Grew up in rural area	−	−

Conclusion

- Patterns in Germany similar to those in the US (Guvenen et al. [2022])
- Focusing on restricted sample hides a large portion of low earners (women and men with lower labor force participation)

- Inequality between high and low earners increases, and is driven by earnings inequality for men
- Inequality patterns over time vary starkly between men and women

Thank you for your attention!

References

Timm Bönke, Giacomo Corneo, and Holger Lüthen. Lifetime earnings inequality in germany. *Journal of Labor Economics*, 33(1):171–208, 2015.

Fatih Guvenen, Fatih Karahan, Serdar Ozkan, and Jae Song. What do data on millions of us workers reveal about lifecycle earnings dynamics? *Econometrica*, 89(5):2303–2339, 2021.

Fatih Guvenen, Greg Kaplan, Jae Song, and Justin Weidner. Lifetime earnings in the united states over six decades. *American Economic Journal: Applied Economics*, 14(4):446–79, 2022.