Ireland in the Global Repository of Income Dynamics: Data Description

Brian E. Higgins
IIES, Stockholm University

Tara McIndoe-Calder
Central Bank of Ireland

Barra Roantree
Trinity College Dublin

Kun Wu*
University College Dublin

December 16, 2024

*Higgins: Institute for International Economic Studies, Stockholm University. Universitetsvägen 10 A, plan 8, Stockholm, SE-106 91. brian.higgins@iies.su.se. McIndoe-Calder: Central Bank of Ireland. New Wapping Street, North Wall Quay, Dublin, D01 F7X3. tara.mcindoecalder@centralbank.ie. Roantree: Trinity College Dublin. Arts Building, College Green, Dublin, D02 PN40. barra.roantree@tcd.ie. Wu: University College Dublin. Geary Institute for Public Policy, Belfield, Dublin, D04 N9Y1. kun.wu1@ucdconnect.ie.

The view expressed here do not necessarily reflect the views of the Central Bank of Ireland nor the Eurosystem of Central Banks. Results are based on analysis of strictly controlled Research Microdata Files provided by the Central Statistics Office (CSO). The CSO does not take any responsibility for the views expressed or the output generated from this research.

1 Data

This document describes the production of the Irish statistics that form part of the Global Repository of Income Dynamics (Guvenen, Pistaferri and Violante, 2022a). To cite the data, please cite Higgins, McIndoe-Calder, Roantree and Wu (2024), which combines this standalone data description with a discussion of the statistics.

1.1 Data Description

Dataset. Our data comes from the Earnings Analysis using Administrative Data Sources (EAADS) dataset, which is an administrative panel dataset held by Ireland's Central Statistics Office (CSO). Earnings information in EAADS is based on tax records collected by the Revenue Commissioners — Ireland's tax authority. Records come from the universal tax return submitted by all registered employers. Before 2018, this was the P35, an end-of-tax-year annual return form, while since 2019, records are based on real-time payroll reports for each payslip.

Timeframe. The data are available from 2011 to 2022.

Earnings Definition. Our measure of earnings is annual individual earnings before tax. This includes taxable bonuses and benefits-in-kind, and excludes pensions and severance payments.

Sampling Frame. The EAADS dataset is a 10 percent sample of employees, randomly sampled based on a digit from the unique personal public service (PPS) number. For each included employee we observe their entire employment history and observe earnings from all employments. The sample size ranges approximately from 170,000 to 230,000 individuals per year. After applying the sample restrictions required by GRID (described below), the sample is approximately between 120,000 and 160,000 per year.

The sample is top and bottom censored. At the bottom, employees earning less than €500 per annum, employments with duration less than two weeks, and secondary employments earning less than €4,000 are excluded. At the top, extremely high earnings values are dropped, as defined by earning that are more than three times the inter-quartile range above the 75th percentile.

The sample includes both public and private sector employment. The sample excludes employments in NACE sectors A (Agriculture, forestry and fishing), T (Activities of households as employers) and U (Activities of extraterritorial organisations and bodies). In line with Eurostat requirements relating to Structure of Earnings Statistics the data used for this analysis has been restricted to employments that were active in the month of October.

GRID Covariates and Other Observables. We use (5 year) age groups and gender as covariates to calculate residual earnings and for sub-group analysis. Unlike some of the other GRID countries, we do not observe education. We also observe employees' nationality and the county of

residence of employers.

1.2 GRID Sample Restrictions

The previous description applies to all data available in the EAADS dataset. Next we describe our variable definitions and additional sample restrictions, where we do our best to be consistent with the other GRID countries, as described by Guvenen, Pistaferri and Violante (2022b).

We impose the same two sample restrictions imposed by other GRID countries. First, we focus on workers between 25 and 55 years old, a range within which most education choices are usually completed and after which workers tend to leave the labor force for retirement. Second, for most of the analysis we drop observations with earnings (defined next) below a threshold (call it $\underline{y}$) to avoid using records from workers without a meaningful attachment to the labor force or with very low earnings, which could skew log-based statistics. Specifically, we discard observations with earnings below what workers would earn if they were to work part-time for one quarter at the Irish national minimum wage.[1]

We construct three separate samples to be used for different parts of the analysis:

1. **The cross-sectional (CS) sample**: This sample is used to compute cross-sectional inequality statistics. All individuals who satisfy the criteria above are included at date t. This is the most comprehensive sample, utilizing the longest possible time series available.

2. **The longitudinal (LX) sample**: This sample is used to study the distribution of earnings changes. It includes all individuals in the CS sample who also have 1-year and 5-year forward earnings changes.[2]

3. **The heterogeneity (H) sample**: This sample examines variation across demographic groups defined by observable characteristics such as age, gender, and permanent income. It includes individuals in the LX sample who also have a permanent earnings measure (definition provided below). We pool observations across years for this analysis.

1.3 Variable Definitions

We construct the following measures of earnings for worker i in year t:

[1] Unlike countries in the first edition of GRID where labor income is top-coded (Brazil, Italy, and Mexico), we do not impute values of top-censored earnings. Our data is top-censored (i.e. set to missing) and not top-coded — Brazil featured similar top-censoring. We have requested more information from the CSO about the extent of top-censoring and will provide that information in future iterations of the data description.

[2] When reporting results that only depend on 1-year changes, we do not impose that individuals have 5-year forward changes. Thus, technically we could refer to these as separate LX^1 and LX^5 samples.

1. Raw real earnings in levels, y_{it}, and logs, $\log(y_{it})$. Real earnings are computed from nominal earnings and deflated to 2018 euros using Irish CPI provided by the OECD.

2. Residualized log earnings, ϵ_{it}. This measure is the residual from a regression of log real earnings on a full set of age dummies, separately for each year and gender. It is intended to control for predictable changes in individual earnings (life cycle and business cycle effects).

3. Permanent earnings, P_{it-1}. They are defined as average earnings over the previous 3 years, $P_{it-1} = \sum_{s=t-2}^{t} y_{is}/3$, where y_{is} can include earnings below $\underline{y}$ for at most 1 year. The measure is intended to average over transitory income changes and proxy for skill levels.

4. 1-year change in residualized log earnings, g_{it}^1. It is the 1-year forward change in ϵ_{it}, defined as $g_{it}^1 = \Delta\epsilon_{it} = \epsilon_{it+1} - \epsilon_{it}$, where earnings must be above $\underline{y}$ for both years.[3]

5. 5-year change in residualized log earnings, g_{it}^5. It is the 5-year forward change in ϵ_{it}, defined as $g_{it}^5 = \Delta\epsilon_{it} = \epsilon_{it+5} - \epsilon_{it}$, where earnings must be above $\underline{y}$ for both years.

1.4 Representativeness

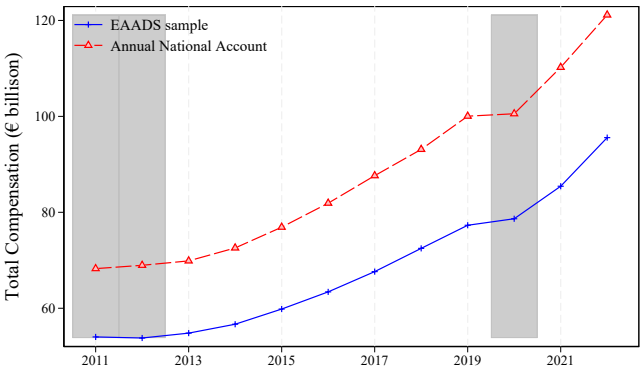
In this section we show that the EAADS dataset is broadly representative of total employment and earnings. Figure 1 compares the time series for total earnings and total employment in national accounts with our estimate from aggregating EAADS. Relative to national accounts, the EAADS dataset captures 90 percent of total employment and 79 percent of total earnings in 2011. We can identify some differences in the sampling frame and definitions of earnings that may explain these discrepancies. In terms of employment, EAADS excludes (i) employment in agriculture, (ii) the top and bottom censoring noted above, and (iii) seasonal employment. The missing seasonal employment occurs because — in line with Eurostat standards — EAADS only includes employees working in October, and thus excludes any seasonal employments taking place during summer or over Christmas. In terms of earnings, the national accounts uses a much broader definition of earnings. In addition to the income included in EAADES, the national accounts include employers' social insurance contributions (PRSI), defined contribution pension contributions actually made by employers, as well as imputed contributions for defined benefit schemes (including public sector schemes). Given the substantial differences in these definitions of earnings, we consider the 80 percent coverage rate of EAADS to be reasonably high.

While there are some differences in the *level* of coverage between EAADS and national accounts, the *trends* are almost identical. The share of employment (90 percent) and earnings (79 percent)

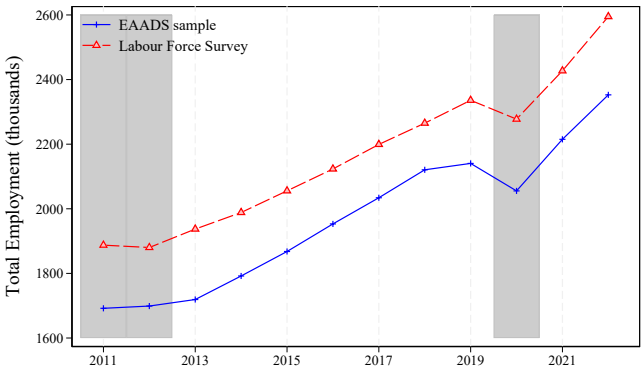
[3] As noted by Guvenen, Pistaferri and Violante (2022b), using "leads" to avoid mechanical mean reversion when conditioning on permanent earnings. In other words, for a given year t, the statistics we are interested in are calculated using t-years forward, whereas the permanent earnings measure which the statistics are conditioned on is computed for years $t-1$ and earlier, avoiding any overlapping years.

Figure 1: Comparison of EAADS with Aggregate Statistics

(a) Total Employment



(b) Total Earnings



Notes.

that are covered are identical at the beginning and end of the sample. Even during the fall in earnings that occurred during the 2020 Covid recession, EAADS captures the same share of employment and marginally less earnings (78 percent). This suggests that the trends identified in the GRID statistics are not affected by differential coverage over time.

References

Guvenen, Fatih, Luigi Pistaferri, and Giovanni L Violante, "The Global Repository of Income Dynamics," *https://www.grid-database.org*, 2022.

_ , _ , and _ , "Global trends in income inequality and income dynamics: New insights from GRID," *Quantitative Economics*, 2022, *13* (4), 1321–1360.

Higgins, Brian E., Tara McIndoe-Calder, Barra Roantree, and Kun Wu, "Inequality and Income Dynamics in Ireland," *Working Paper*, 2024.