# Source Data Reproduction Attempt to Evaluate a Claim from Horvat\_EurSocioRev\_2011\_lxXV

**Reproduction analyst(s):** Nate Breznau,

University of Bremen, [breznau.nate@gmail.com](mailto:breznau.nate@gmail.com)

**SCORE RR ID:** y791

**OSF Project:** <https://osf.io/q5szk>

## Transparency Trail

**Data acquisition:** Initial contact with the first listed author of the original study (Horvat and Evans 2011) (28/9/2020) redirected me to the second listed author (also 28/9/2020) who pointed out that the first wave of data should be available on ESRC’s data portal, but that the ‘second wave may or may not be there’. I confirmed that the first wave was there, but the second wave was not. There were some back and forth between the second author and further potential data holders, and eventually a third-party was found to be the holder of the entire datafile in Stata format. The data come from two independent studies that had similar questions and sampling strategies, therefore they are sufficiently similar to constitute a single ‘survey’ with two waves. Following the convention in the code book provided by the third-party contact these data will be labeled “EUREQUAL 1993-2007 Survey”, EUREQUAL for short.

The EUREQUAL survey aimed to attain for Central-/Eastern-Europe and Russia “national probability samples employing standardized rules for respondent selection procedures, with sampling frames designed to ensure their representativeness in each country. Extensive measures were taken to ensure reliability and cross-national comparability of questionnaires, including pilot-testing, translation and back-translation” (Horvat and Evans 2011:712). Although the replicator has no way to check these methods, the description sounds consistent with high standards in survey research.

**Data analysis:** In this reproduction I acted as both data finder and data analyst. Thus, no special alterations were necessary between acquisition and analysis, other than using the full dataset rather than a 5% random sample. I used R Studio, my IDE of choice, to run R statistical software and to install the ‘MASS’ package (Venables and Ripley 2002) because it has a cumulative link model function which is normally analogous to an “ordered probit” – the modelling strategy reported by Horvat and Evans (2012:717).

Upon close inspection of the data and the codebook, it became clear that the data were not identical to those reportedly used in the original study. This conclusion is based on some problems in preparing the variables for analysis. In particular, data for Russia in 1996 does not appear to be recoded into ISCED-97, or if so, it contains many additional categories or the wrong numerical representations of the ISCED scale (‘std\_education’ variable label in original data). I checked the country-specific education variable (v198) in the data and the codebook, but the values simply do not match what values exist in the data and what are reported in the codebook for Russia. I checked further Belarus and Bulgaria on the country-specific education variable and Bulgaria contains some values that are not reported in the codebook, so hand-coding of Russia (or any other country) into ISCED seemed unwise. Therefore, I simply took the purported ISCED-97 values and recoded Russian education scores from 7-12 into “high” education, as the original study collapsed ISCED-97 into “low”, “middle” and “high” education.

It also became apparent that the data are different because of variation in case numbers and descriptive statistics. I attempted to reproduce Tables 1-3 from the original manuscript before trying to reproduce the analyses. Whether the data are in fact different or the authors somehow manipulated the data in a way other than I leading to variation in descriptive outcomes, I cannot conclude. The statistical differences are presented in my own Tables 1-3 in Appendix. For example, the first row of Table 1R my own replication, shows that the percentage of the sample that are aged 18-29 that have a “low” level of education are 29%; however, Table 1O their original Table 1, suggests it is 13%. This is not a product of weighting as I produced Table 1R with and without weights. This is a large discrepancy and may be seen in several of the descriptive means when comparing these two tables, and Tables 2R and 2O and 3R and 3O.

Link to analysis script(s): <https://osf.io/q5szk/>, see GitHub sub-folder “nbreznau/score\_horvath”

### Analysis attempt 1

Table 4 is the first analysis performed in the reproduction effort. It is ostensibly the same analysis performed by the original study. At first, I thought the results might differ due to slight differences in the data used. In particular, the final version of the data provided by the original authors appears to have slightly different descriptive statistics. Further communication with the original authors was not undertaken at this point given the goals of testing the independent reproducibility of the research.

**Table 4. Reproduction of Horvat and Evans (2011:Table 5) cumulative link models (ordered probit) of subjective household standard of living over the past five years.**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable / Parameter | Model 1 | Model 2 | Model 3 |
| Age 30-44 | -0.41 \*\*\* | -0.46 \*\*\* | -0.46 \*\*\* |
| Age 44-59 | -0.81 \*\*\* | -0.82 \*\*\* | -0.82 \*\*\* |
| Age >60 | -0.97 \*\*\* | -0.82 \*\*\* | -0.75 \*\*\* |
| Year 2007 | 0.75 \*\*\* | 0.73 \*\*\* | 0.75 \*\*\* |
| Age 30-44\*year'07 | 0.00 | 0.02 | -0.02 |
| Age 44-59\*year'07 | -0.11 | -0.10 | -0.11 |
| Age >60\*year'07 | -0.38 \*\*\* | -0.40 \*\*\* | -0.39 \*\*\* |
| Female |  | -0.15 \*\*\* | -0.10 \*\*\* |
| Educ mid |  | 0.08 \*\* | 0.02 |
| Educ high |  | 0.21 \*\*\* | 0.17 \*\*\* |
| EGP: routine non-man |  | -0.08 \* | -0.06 |
| EGP: Self |  | 0.22 \*\*\* | 0.26 \*\*\* |
| EGP: Skilled |  | -0.16 \*\*\* | -0.14 \*\*\* |
| EGP: Unskilled |  | -0.17 \*\*\* | -0.13 \*\*\* |
| EGP: Farmers |  | -0.18 \*\*\* | -0.14 \*\*\* |
| EGP: Never had a job |  | -0.32 \*\*\* | -0.19 \*\*\* |
| Income mid |  | 0.14 \*\*\* | 0.11 \*\*\* |
| Income high |  | 0.41 \*\*\* | 0.36 \*\*\* |
| Income missing |  | 0.21 \* | 0.18 |
| Pensions and benefits |  |  | -0.09 \*\* |
| Unemployed |  |  | -0.20 \*\*\* |
| Car |  |  | 0.27 \*\*\* |
| Country-specific intercepts | Yes | Yes | Yes |
| k1 | -2.51 \*\*\* | -2.45 \*\*\* | -2.21 \*\*\* |
| k2 | -1.61 \*\*\* | -1.54 \*\*\* | -1.29 \*\*\* |
| k3 | 0.59 \*\*\* | 0.68 \*\*\* | 0.99 \*\*\* |
| k4 | 3.18 \*\*\* | 3.30 \*\*\* | 3.64 \*\*\* |
| Observations | 35648 | 35648 | 35648 |
| R2 Nagelkerke | 0.068 | 0.086 | 0.120 |
| log-Likelihood | -45584.674 | -45258.680 | -44644.165 |
| *\* p<0.05   \*\* p<0.01   \*\*\* p<0.001* | | | | |

Note: un-exponentiated coefficients presented, country intercepts omitted to save space (see Appendix for full table). Analysis done with ‘polr’ function from package ‘MASS’.

### Analysis attempt 2

I suspected that the original authors used Stata or SPSS software, as R was virtually non-existent among social scientists 15 years ago when the did the analysis. Therefore, I also attempted to reproduce the results using Stata and the ‘oprobit’ function to compare with my results from R and the package ‘MASS’ with the ‘polr’ function. Table 5 reports the results which are much more closely in-line with the original results.

**Table 5. Reproduction of Horvat and Evans (2011:Table 5) cumulative link models (ordered probit) of subjective household standard of living over the past five years.**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable / Parameter | Model 1 | Model 2 | Model 3 |
| Age 30-44 | -0.238\*\*\* | -0.265\*\*\* | -0.282\*\*\* |
| Age 44-59 | -0.467\*\*\* | -0.473\*\*\* | -0.481\*\*\* |
| Age >60 | -0.568\*\*\* | -0.478\*\*\* | -0.437\*\*\* |
| Year 2007 | 0.444\*\*\* | 0.433\*\*\* | 0.402\*\*\* |
| Age 30-44\*year'07 | -0.001 | 0.007 | 0.011 |
| Age 44-59\*year'07 | -0.077\* | -0.068 | -0.055 |
| Age >60\*year'07 | -0.225\*\*\* | -0.232\*\*\* | -0.201\*\*\* |
| Female | -0.013\*\*\* | -0.012\*\*\* | -0.010\*\*\* |
| Educ mid |  | -0.084\*\*\* | -0.068\*\*\* |
| Educ high |  | 0.043\*\* | 0.030\* |
| EGP: routine non-man |  | 0.101\*\*\* | 0.081\*\*\* |
| EGP: Self |  | -0.050\* | -0.045\* |
| EGP: Skilled |  | 0.125\*\*\* | 0.108\*\* |
| EGP: Unskilled |  | -0.100\*\*\* | -0.083\*\*\* |
| EGP: Farmers |  | -0.115\*\*\* | -0.093\*\*\* |
| EGP: Never had a job |  | -0.127\*\*\* | -0.102\*\*\* |
| Income mid |  | -0.192\*\*\* | -0.142\*\*\* |
| Income high |  | 0.084\*\*\* | 0.056\*\*\* |
| Income missing |  | 0.250\*\*\* | 0.196\*\*\* |
| Pensions and benefits |  | 0.133\* | 0.107 |
| Unemployed |  |  | -0.068\*\*\* |
| Car |  |  | -0.131\*\*\* |
| Country-specific intercepts | Yes | Yes | Yes |
| k1 | -1.575\*\*\* | -1.535\*\*\* | -1.497\*\*\* |
| k2 | -1.069\*\*\* | -1.026\*\*\* | -0.986\*\*\* |
| k3 | 0.271\*\*\* | 0.328\*\*\* | 0.374\*\*\* |
| k4 | 1.624\*\*\* | 1.703\*\*\* | 1.755\*\*\* |
| Observations | 35648 | 35648 | 35648 |
| R2 Nagelkerke | 0.026 | 0.034 | 0.037 |
| log-Likelihood | -45500 | -45200 | -45000 |
| *\* p<0.05   \*\* p<0.01   \*\*\* p<0.001* | | | | |

Note: un-exponentiated coefficients presented, country intercepts omitted to save space (see Appendix for full table). Analysis done with ‘oprobit’ function from Stata v15.

## Claim evaluation

### Single-trace claim

**Coded claim 4 text (original paper):** “The more polarized views, with which the elderly and the young have come to describe their prospective economic situation, cannot be explained away by changes in socio-demographic characteristics and resources. The introduction of socio-demographic controls to Model 2 and resources to Model 3 has no significant effect on the magnitude of the interaction term (from Model 3 in Table 5, ‘>60 x year’07’ term: estimate = -0.22; SE = 0.04; P < 0.001).”

**Reproduction data source(s):**

<https://osf.io/xjz7p/>

**Description of reproduction data:**

EUREQUAL as described in the section above “Transparency Trail” is available in a single Stata (version 13+) datafile “Mass\_Public\_Surveys\_1993-2007.dta” available in the “data” folder of the GitHub plugged-in OSF workflow.

**Primary reproduction criteria**

| Criterion | Original value | Precise reproduction | Approximate reproduction | Non-reproduction | **Reproduction result** |
| --- | --- | --- | --- | --- | --- |
| Sample size | 32,417 | 32,417 | 27,554 < x < 37,280 | x < 27,554 or x > 37,280 | **Approximate reproduction** |
| Focal coefficient | probit regression coefficient = -0.22 | probit regression coefficient = -0.22 | -0.253 < x < -0.187 | x < -0.253 or x > -0.187 | **Approximate reproduction** |
| Focal test statistic | NA | NA | NA | NA | **NA** |
| Focal effect size | NA | NA | NA | NA | **NA** |
| Focal p-value | p < 0.001 | p < 0.001 | p < 0.051 | p > 0.051 | **Precise reproduction** |

**Analyst success criteria:** NA

**Reproduction outcome:** Based on these criteria, the claim reproduced.

**Discussion:** The main coefficient is approximately reproduced. It is very close and would lead to the same conclusions as in the original study. Thus, as long as one relies on the results from Table 5 (Stat v. R), then it is safe to say any differences are negligible. Most likely the data used 15 years ago, and the final version of the EUREQUAL data prepared to be shared publicly had slight variations in either coding or cases. Discovering these exact reasons for variation is beyond the scope of this study.

## General discussion (optional)

What is fascinating is that the claim only reproduced in Stata. Using a standard ordered probit model in R led to slightly different results. This is in itself a finding worthy of further consideration, but beyond the scope of this reproduction attempt. If I was not reasonably familiar with Stata and only used R, the conclusion would be different. Namely the effect of interest would have fallen outside the focal coefficient cut-off range and led to a non-reproduction conclusion.

## Description of materials provided

The entire workflow can be followed in the Appendix which are Markdown flies that were knitted together. The final Table 5 was produced with the Stata file not part of the markdown workflow. As it was unexpected that Stata and R would differ, the Stata portion of the project was added only at the last minute.

Files contained in the GitHub repository stored within the OSF project:

|  |  |
| --- | --- |
| File Name | Description |
| O1\_Data\_Prep.Rmd | This script imports the original data, engages in recoding and reproduces Tables 1-3. |
| 02\_Pre\_Analysis.Rmd | This prepares the main three models and randomly selects 5% of the sample to test that they do run. |
| 03\_Main\_Analysis.Rmd | This runs the models on the full dataset and compiles the results. It also compares the results from Stata and R. |
| 03\_Main\_Analysis\_Stata.do | This runs the main models in Stata and provides the results for Table 5 in this document which is the reproduction of choice. |

## 

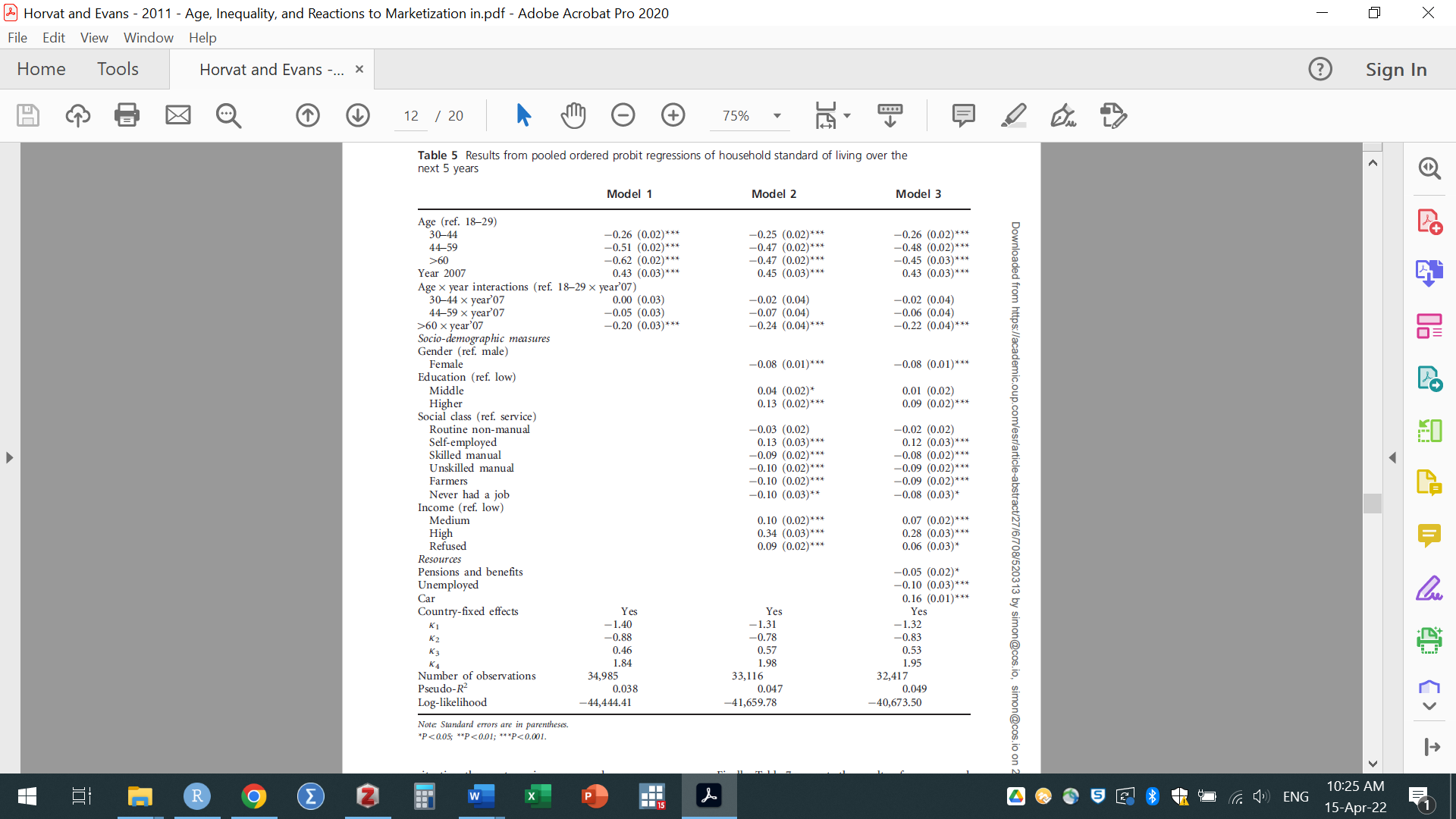
## References

Horvat, P., and G. Evans. 2011. “Age, Inequality, and Reactions to Marketization in Post-Communist Central and Eastern Europe.” *European Sociological Review* 27(6doi: 10.1093/esr/jcq033):708–27.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer.

## Appendix

Note that the workflow documents were converted to PDF from HTML format. The tables say ‘log-odds’ but they are actually logistic coefficients. I did not have time before the Apr-15 deadline to fix this.



Note: Table 5 above is a snapshot (image file) from the original PDF.