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

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.

Table 5 Results from pooled ordered probit regressions of household standard of living over the next 5 years

	Model 1	Model 2	Model 3
Age (ref. 18–29)			
30–44	−0.26 (0.02)***	−0.25 (0.02)***	−0.26 (0.02)***
44–59	−0.51 (0.02)***	−0.47 (0.02)***	−0.48 (0.02)***
>60	−0.62 (0.02)***	−0.47 (0.02)***	−0.45 (0.03)***
Year 2007	0.43 (0.03)***	0.45 (0.03)***	0.43 (0.03)***
Age × year interactions (ref. 18–29 × year'07)			
30–44 × year'07	0.00 (0.03)	−0.02 (0.04)	−0.02 (0.04)
44–59 × year'07	−0.05 (0.03)	−0.07 (0.04)	−0.06 (0.04)
>60 × year'07	−0.20 (0.03)***	−0.24 (0.04)***	−0.22 (0.04)***
Socio-demographic measures			
Gender (ref. male)			
Female		−0.08 (0.01)***	−0.08 (0.01)***
Education (ref. low)			
Middle		0.04 (0.02)*	0.01 (0.02)
Higher		0.13 (0.02)***	0.09 (0.02)***
Social class (ref. service)			
Routine non-manual		−0.03 (0.02)	−0.02 (0.02)
Self-employed		0.13 (0.03)***	0.12 (0.03)***
Skilled manual		−0.09 (0.02)***	−0.08 (0.02)***
Unskilled manual		−0.10 (0.02)***	−0.09 (0.02)***
Farmers		−0.10 (0.02)***	−0.09 (0.02)***
Never had a job		−0.10 (0.03)**	−0.08 (0.03)*
Income (ref. low)			
Medium		0.10 (0.02)***	0.07 (0.02)***
High		0.34 (0.03)***	0.28 (0.03)***
Refused		0.09 (0.02)***	0.06 (0.03)*
Resources			
Pensions and benefits			−0.05 (0.02)*
Unemployed			−0.10 (0.03)***
Car			0.16 (0.01)***
Country-fixed effects	Yes	Yes	Yes
κ_1	−1.40	−1.31	−1.32
κ_2	−0.88	−0.78	−0.83
κ_3	0.46	0.57	0.53
κ_4	1.84	1.98	1.95
Number of observations	34,985	33,116	32,417
Pseudo-R^2	0.038	0.047	0.049
Log-likelihood	−44,444.41	−41,659.78	−40,673.50

Note: *Standard errors are in parentheses.*
*$P<0.05$; **$P<0.01$; ***$P<0.001$.

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

Tech 1. Data Prep.

This workflow file takes the source data, recodes variables and prepares descriptive tables for comparison.

The original Tables 1-3 were extracted from the original study PDF using Adobe Acrobat 'copy with formatting' function and placed into the document 'orig_tables.xlsx'.

```
pacman::p_load("tidyverse","readxl","foreign","skimr","fastDummies","kableExtra","MASS","sjPlot","webshot","knitr")
```

Import Data

Datafile acquired after emails to Horvat then to Evans then to the ultimate source of this aggregate file "Mass_Public_Surveys_1993-2007.dta" with Ksenia Northmore-Ball (Nov. 8th, 2020)

```
# original data
df_origa <- readstata13::read.dta13("data/Mass_Public_Surveys_1993-2007.dta")
df_orig_labels <- readstata13::read.dta13("data/Mass_Public_Surveys_1993-2007.dta", generate.factors = T)

# original tables 1 & 2 from the article
Tbl1_orig <- read_xlsx("data/orig_tables.xlsx", sheet = "Tbl1", col_names = F)
Tbl2_orig <- read_xlsx("data/orig_tables.xlsx", sheet = "Tbl2", col_names = F)
Tbl3_orig <- read_xlsx("data/orig_tables.xlsx", sheet = "Tbl3", col_names = F)

# codebook
codebook <- read.csv("data/codebook.csv", header = T)
```

v270 Main source of income	Code
Earnings from employment (own or partner's)	1
Pensions and benefits	2
Student stipend	3
Other state benefit	4
Interest from savings or property	5
Dependent on family/relatives	6

v200 Work Status	Code
in paid work (including self-employment)	1
full-time student	2
in military service	3
unemployed	4

permanently sick or disabled	5
completely retired from work	6
looking after the home	7
other	8
NA	9
demobilized	10
vacation without salary	11

Education = “Education is measured by three categories: low, middle, and higher education. Low education means no educational qualifications beyond the compulsory level. Middle education corresponds to completed secondary education and higher education corresponds to completed further or university education.” (p. 713)

ISCED-97 0 pre-primary 1 primary/1st stage basic ed 2 lower secondary 3 upper secondary 4 post-secondary non-tertiary 5 1st stage tertiary 6 2nd stage tert

v269 income per month Income recoded into terciles by country-year (see footnote 3)

Class

“Social class is measured by a six-category version of the Erikson–Goldthorpe class schema (Erikson and Goldthorpe, 1992), based on occupational measures of class position: service class, routine non-manual workers, self-employed workers, manual supervisors and skilled manual workers, semi-skilled and unskilled manual workers, and farmers or agricultural workers. A residual category ‘never had a paying job’ denotes respondents whose social class was ambiguous or missing but who reported never having been in paid employment elsewhere in the survey. Women with missing social class data were classified according to their husband’s class. Previous research on Eastern Europe suggests that occupational measures of class position perform adequately in the Eastern European context and successfully differentiate individuals in terms of their level of income, their degree of economic security, and chances of economic advancement (Evans, 1997; Evans and Mills, 1999).”

Students dropped

Age cat_age: cross-temporally and cross-nationally consistent age categories: 1: -29; 2: 30-44; 3: 45-59; 4: 60+

Recode Analysis Variables

```
df_orig <- df_origa %>%
  mutate(stdliv_past5 = car::recode(v272, "8 = 3"), # don't knows were recoded into middle cat
 in original study
         stdliv_past5_di = car::recode(stdliv_past5, "c(1,2) = 1; c(4,5) = 2; c(3) = 3"),
         stdliv_next5 = car::recode(v273, "8 = 3"),
         stdliv_next5_di = car::recode(stdliv_next5, "c(1,2) = 1; c(4,5) = 2; c(3) = 3"),
         noway_future_improve = as.numeric(as.character(ifelse(v276 == "no any way", 1, 0))),
         mkt_econ_eval = car::recode(v4, "'dont know' = 3"),
         mkt_econ_eval_di = car::recode(mkt_econ_eval, "c('very positively','positively') = 1;
 c('negatively','very negatively') = 2; c('neither positively nor negatively') = 3"),
         student = ifelse(v200 == 2 | v200 == 9, NA, 0), # remove students and NA's
         wave = as.numeric(car::recode(year, "c('1993','1994','1995','1996') = '1993'; c('2007
') = '2007'; c('1997','1998','2001','2002','2003','2004','2005','2006') = NA")),
         female = as.numeric(v298) - 1,
```

```
        pensions = ifelse(is.na(v270), NA, ifelse(v270==2, 1, 0)),
        unemployed = ifelse(v200 == 8, NA, ifelse(v200==4, 1, 0)),
        car_owner = ifelse(v262 ==1, 1, 0),
        education = car::recode(std_education, "c(0,1,2) = 1; c(3,4) = 2; c(5,6) = 3; c(99) =
 NA"), # some had 99 in the std_education variable still, made into primary
        education_a = car::recode(std_education, "c(1,2,95,96) = 1; c(3,4,5,6,8) = 2; c(7,9,1
0,11,12) = 3; c(98,99,14) = NA"), # 4 and 5 are questionable categories here
        education = ifelse(cntry == "russia" & year == 1996, education_a, education),
        # There is a problem with Russia in 1996, seems that it was not recoded into ISCED
        EGP6 = car::recode(rclass10, "'Missing in 93-03 data' = 8;'semi-unskilld manual' = 5;
 'skilled manual' = 4; 'higher controllers' = 1;'lo controllers' = 1;  'routine nonmanual'= 2;
 'sempl without empl' = 3; 'sempl with emp' = 4; 'selfempl farm' = 6; 'farm labor'=6; 'manual
supervisor' = 4; 'Missing Occupation Code 07' = 8; 'Not ISKO codes' = 8; 'ISKO Coded - no matc
h' = 8; 'Never had a paying job' = 7"),
        EGP6 = ifelse(EGP6 == 8 & r2class10 == "never had a paying job", "7", EGP6),
        EGP6 = ifelse(is.na(v201), EGP6, ifelse(v201 == "no", "7", EGP6)), # never had a paid
 job
        EGP6 = ifelse(EGP6 == 8, NA, EGP6)
        ) %>%
  group_by(cntry, year) %>%
  mutate(income = ifelse(is.na(v269), 4, ntile(v269, 3))) %>%
  ungroup()
```

Weights

There are some curiosities with the weights. There are no weights for 1993 and other weights are NA meaning that full weighting is not possible. I recode NA to 1 for now to simply preserve cases.

```
# some weights are NA, replace with 1 (actually there are no weights for 1993, so this is not
really helpful)
df_orig$weight <- ifelse(is.na(df_orig$v461), 1, df_orig$v461)
```

Complete Dataframe

And generate group descriptive means

```
# create complete cases df
df_orig_complete <- df_orig[!is.na(df_orig["education"]),]
# df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["income"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["EGP6"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["female"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["wave"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["student"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["pensions"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["unemployed"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["car_owner"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["stdliv_past5"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["stdliv_next5"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["mkt_econ_eval"]),]
df_orig_complete <- df_orig_complete[!is.na(df_orig_complete["noway_future_improve"]),]
# create factor dummies
```

```
df_orig_complete <- dummy_cols(df_orig_complete, select_columns = c("EGP6", "income", "educati
on","stdliv_past5_di","stdliv_next5_di","mkt_econ_eval_di"))

#create new group ID
df_orig_complete$group <- df_orig_complete$cat_age + (100*df_orig_complete$wave)

# cases per group

cases <- df_orig_complete %>%
  group_by(group) %>%
  count() %>%
  ungroup() %>%
  dplyr::select(-group) %>%
  t()

  #df_orig <- select(df_orig, v201, rclass10, EGP6, everything())
```

Replicate Table 1

```
# get weighted means by group
Tbl1_rep <- apply(df_orig_complete[,c("education_1","education_2","education_3","income_1","in
come_2","income_3","EGP6_1","EGP6_2","EGP6_3","EGP6_4","EGP6_5","EGP6_6","EGP6_7","female")],
2, function(x) {sapply(split(data.frame(df_orig_complete[,"weight"], x), df_orig_complete$grou
p), function(y) weighted.mean(y[,2], w = y$weight))})

Tbl1_rep <- as.data.frame(t(round(Tbl1_rep*100, 0)))

Tbl1_rep_unw <- apply(df_orig_complete[,c("education_1","education_2","education_3","income_1"
,"income_2","income_3","EGP6_1","EGP6_2","EGP6_3","EGP6_4","EGP6_5","EGP6_6","EGP6_7","female"
)], 2, function(x) {sapply(split(data.frame(df_orig_complete[,"weight"], x), df_orig_complete$
group), function(y) mean(y[,2], w = y$weight))})

Tbl1_rep_unw <- as.data.frame(t(round(Tbl1_rep_unw*100, 0)))

Tbl1_rep[nrow(Tbl1_rep)+1,] <- cases

Tbl1_rep_unw[nrow(Tbl1_rep_unw)+1,] <- cases

rownames(Tbl1_rep)[15] <- "N"
rownames(Tbl1_rep_unw)[15] <- "N"

# create csv for easy importing in final report
write.csv(Tbl1_rep, here::here("results", "Tbl1_rep_unw.csv"), row.names = F)

Tbl1_rep_out <- kable_styling(kable(Tbl1_rep, col.names = c("Wave 1, Age 18-29","Wave 1, Age 3
0-49","Wave 1, Age 50-59","Wave 1, Age 60+","Wave 2, Age 18-29","Wave 2, Age 30-49","Wave 2, A
ge 50-59","Wave 2, Age 60+"), caption = "Table 1R. Replicated Weighted. Age and socio-demograp
hic outcomes weighted in percentages by age group"))
save_kable(Tbl1_rep_out, file = "results/Tbl1_rep.htm")

Tbl1_rep_out_unw <- kable_styling(kable(Tbl1_rep, col.names = c("Wave 1, Age 18-29","Wave 1, A
ge 30-49","Wave 1, Age 50-59","Wave 1, Age 60+","Wave 2, Age 18-29","Wave 2, Age 30-49","Wave
```

```
2, Age 50-59","Wave 2, Age 60+"), caption = "Table 1R. Replicated Unweighted. Age and socio-de
mographic outcomes weighted in percentages by age group"))


#webshot("results/Tbl1_rep.htm", file = "results/Tbl1_rep.png")
```

Table 1R Weighted

```
Tbl1_rep_out
```

Table 1R. Replicated Weighted. Age and socio-demographic outcomes weighted in percentages by age group

	Wave 1, Age 18-29	Wave 1, Age 30-49	Wave 1, Age 50-59	Wave 1, Age 60+	Wave 2, Age 18-29	Wave 2, Age 30-49	Wave 2, Age 50-59	Wave 2, Age 60+
education_1	29	26	39	62	26	42	38	72
education_2	54	48	39	25	24	35	48	15
education_3	17	25	22	13	50	23	15	12
income_1	26	24	30	56	16	32	38	39
income_2	34	35	34	31	27	31	42	45
income_3	39	41	36	13	57	37	20	15
EGP6_1	22	31	32	24	14	23	18	24
EGP6_2	15	13	11	10	7	22	15	13
EGP6_3	4	3	2	1	2	4	14	0
EGP6_4	23	22	21	18	55	26	16	19
EGP6_5	18	20	20	21	9	17	32	31
EGP6_6	8	10	12	22	0	4	4	10
EGP6_7	10	1	1	5	11	4	2	3
female	52	52	54	57	38	66	65	58
N	3827	7325	5766	5226	2086	3596	3876	3946

Table 1R Unweighted

```
Tbl1_rep_out_unw
```

Table 1R. Replicated Unweighted. Age and socio-demographic outcomes weighted in percentages by age group

	Wave 1, Age 18-29	Wave 1, Age 30-49	Wave 1, Age 50-59	Wave 1, Age 60+	Wave 2, Age 18-29	Wave 2, Age 30-49	Wave 2, Age 50-59	Wave 2, Age 60+
education_1	29	26	39	62	26	42	38	72
education_2	54	48	39	25	24	35	48	15
education_3	17	25	22	13	50	23	15	12
income_1	26	24	30	56	16	32	38	39
income_2	34	35	34	31	27	31	42	45
income_3	39	41	36	13	57	37	20	15
EGP6_1	22	31	32	24	14	23	18	24
EGP6_2	15	13	11	10	7	22	15	13
EGP6_3	4	3	2	1	2	4	14	0
EGP6_4	23	22	21	18	55	26	16	19
EGP6_5	18	20	20	21	9	17	32	31
EGP6_6	8	10	12	22	0	4	4	10
EGP6_7	10	1	1	5	11	4	2	3
female	52	52	54	57	38	66	65	58
N	3827	7325	5766	5226	2086	3596	3876	3946

Table 1 Original

```
Tbl1_orig <- Tbl1_orig[-c(1:3),]
colnames(Tbl1_orig) <- c("Variable","y93-96 a18-29","y93-96 a30-44","y93-96 a 45-59","y93-96 a
60+","y07 18-29","y07 30-44","y07 45-59","y07 60+")

Tbl1_orig[,2:9] <- lapply(Tbl1_orig[,2:9], function (x) ifelse(is.na(x), "", x))

Tbl2_orig <- Tbl2_orig[-c(1:2),]

colnames(Tbl2_orig) <- c("Variable","y93-96 a18-29","y93-96 a30-44","y93-96 a 45-59","y93-96 a
60+","y07 18-29","y07 30-44","y07 45-59","y07 60+")

Tbl2_orig[,2:9] <- lapply(Tbl2_orig[,2:9], function (x) ifelse(is.na(x), "", x))

Tbl1_orig$Variable <- c(NA, "Education_1_Low", "Education_2_Mid", "Education_3_Hi", NA, "Incom
```

```
e_1_Low", "Income_2_Mid", "Income_3_Hi", NA, "EGP6_1_Service", "EGP6_2_Rtn_NonMan", "EGP6_3_Se
lfEmp", "EGP6_4_SkilMan", "EGP6_5_Unskilled", "EGP6_6_Farmers", "EGP6_7_NeverHadJob", "Female"
, "N")

Tbl1_orig <- subset(Tbl1_orig, !is.na(Variable))

kable_styling(kable(Tbl1_orig, caption = "Table 1O. Original. Age and socio-demographic outcom
es weighted in percentages by age group"))
```

Table 1O. Original. Age and socio-demographic outcomes weighted in percentages by age group

Variable	y93-96 a18-29	y93-96 a30-44	y93-96 a 45-59	y93-96 a60+	y07 18-29	y07 30-44	y07 45-59	y07 60+
Education_1_Low	13	14	29	56	11	9	13	40
Education_2_Mid	75	66	51	31	66	69	67	45
Education_3_Hi	13	21	20	13	23	22	19	15
Income_1_Low	18	16	20	42	19	14	21	32
Income_2_Mid	53	54	52	49	44	47	50	59
Income_3_Hi	29	31	28	9	37	39	29	9
EGP6_1_Service	21	30	32	23	30	32	29	27
EGP6_2_Rtn_NonMan	16	14	11	10	18	17	14	12
EGP6_3_SelfEmp	6	5	3	2	4	5	5	2
EGP6_4_SkilMan	21	20	20	17	14	17	18	16
EGP6_5_Unskilled	19	20	20	21	18	20	24	24
EGP6_6_Farmers	9	10	13	23	3	5	8	15
EGP6_7_NeverHadJob	9	1	1	4	14	4	2	4
Female	51	52	53	55	54	56	57	59
N	3800	6897	5542	5019	2107	3667	3945	4045

The results do not match very well, especially in the education category. Actually it appears that only 2007 has weights. Running Table 1 without weights changes very little. Therefore, a proper test requires hand coding by country; however, the country-specific education codes in the codebook for v198 are incorrect, e.g., I tested Bulgaria and Belarus 1993. Belarus looks ok except two missing codes are not present, but Bulgaria has several extra codes than what is listed in the codebook provided (values 8-12). Therefore, this hand coding is not possible. Russia has different codes altogether, therefore I had to guess how to code it under the assumption the variable was similarly ordinal.

Education Peculiarities

ISCED in all countries except Russia

```
unique(df_orig$std_education[df_orig$cntry != "russia"])
```

```
## [1]  5  6  3  4  1  0  2 NA 99
```

Russia not in ISCED

Extra categories not in ISCED-97 or in other countries

```
unique(df_orig$std_education[df_orig$cntry == "russia"])
```

```
##  [1]  3  4  2  1 NA  5  0  9 10  7  8  6 11 12 96 95
```

Check original education variable

Russia

Not coded following original codebook (see “data” folder). At least one category is present in data (“14”) that is not in codebook

RUSSIA 1993 V198 highest education level 1 primary school 2 secondary school 3 high school 4 professional courses 5 vocational school 6 technical secondary school 7 vocational post-school 8 technical college 9 incomplete high education(institute,university,academy) 10 high education(institute,university,academy) 11 additional training courses 12 degree 95 no educational qualifications 96 never went to school 98 don’t know

```
unique(df_orig$v198[df_orig$cntry == "russia"])
```

```
##  [1]  3 10  7  6  2  9  8  5  1  4 11 NA 12 96 95 99 14
```

Belarus

BELARUS 1993 V198 education 1 elementary school 2 junior high school 3 high school 4 professional training courses 5 regular factory-and-workshop school,industrial training scho 6 industrial training high school 7 college (for nurses, elementary school teachers, musicians) 8 technical college 9 bachelor’s degree 10 master’s degree (university, academy etc) 11 postgraduate courses 12 Ph. Degree 95 no certificates of education of any kind 96 never went to school 98 don’t know

```
unique(df_orig$v198[df_orig$cntry == "belarus"])
```

```
##  [1]  8 10  3  6  7  4 11  9  5  1 95  2 NA 12
```

Bulgaria

BULGARIA 1993 V198 education 1 NO COMPLETED ED. 2 ELEMENTARY 3 PRIMARY 4 HIGH SCHOOL 5 SECONDARY VOCATIONAL 6 COLLEGE 7 HIGHER 98 DON T KNOW

```
unique(df_orig$v198[df_orig$cntry == "belarus"])
```

```
##  [1]  8 10  3  6  7  4 11  9  5  1 95  2 NA 12
```

Table 2 Replicated

```
# get weighted means by group
Tbl2_rep <- apply(df_orig_complete[,c("pensions","unemployed","car_owner")], 2, function(x) {s
apply(split(data.frame(df_orig_complete[,"weight"], x), df_orig_complete$group), function(y) w
eighted.mean(y[,2], w = y$weight))})

Tbl2_rep <- as.data.frame(t(round(Tbl2_rep, 3)))

Tbl2_rep[nrow(Tbl2_rep)+1,] <- cases

rownames(Tbl2_rep)[4] <- "N"

# create csv for easy importing in final report
write.csv(Tbl2_rep, here::here("results", "Tbl2_rep.csv"), row.names = F)

Tbl2_rep[1,] <- round(as.numeric(Tbl2_rep[1,])*100,0)
Tbl2_rep[2,] <- round(as.numeric(Tbl2_rep[2,])*100,0)
Tbl2_rep[3,] <- round(as.numeric(Tbl2_rep[3,])*100,0)
Tbl2_rep[4,] <- round(as.numeric(Tbl2_rep[4,]),0)

kable_styling(kable(Tbl2_rep, caption = "Table 2R. Replicated. Age-based inequality in resourc
es in percentage", col.names = c("Wave 1, Age 18-29", "Wave 1, Age 30-44", "Wave 1, Age 45-59"
, "Wave 1, 60+", "Wave 2, Age 18-29", "Wave 2, Age 30-44", "Wave 2, Age 45-59", "Wave 2, 60+")
))
```

Table 2R. Replicated. Age-based inequality in resources in percentage

	Wave 1, Age 18-29	Wave 1, Age 30-44	Wave 1, Age 45-59	Wave 1, 60+	Wave 2, Age 18-29	Wave 2, Age 30-44	Wave 2, Age 45-59	Wave 2, 60+
pensions	2	3	22	86	2	5	22	94
unemployed	16	9	6	0	8	11	10	1
car_owner	37	44	40	22	37	53	62	28
N	3827	7325	5766	5226	2086	3596	3876	3946

Table 2. Original

```
kable_styling(kable(Tbl2_orig, caption = "Table 2O. Original. Age-based inequality in resource
s in percentage"))
```

Table 2O. Original. Age-based inequality in resources in percentage

Variable	y93-96 a18-29	y93-96 a30-44	y93-96 a 45-59	y93-96 a60+	y07 18-29	y07 30-44	y07 45-59	y07 60+
Pensions and benefits	2	3	22	86	5	6	21	88
Unemployed	12	10	6	0.4	8	7	8	0.7
Car-ownership	43	47	43	24	58	61	53	28
N	3800	6897	5542	5019	2107	3667	3945	4045

Table 3R. Replicated Weights.

```
Tbl3_rep <- apply(df_orig_complete[,c("stdliv_past5_di_1","stdliv_past5_di_2","stdliv_next5_di_1","stdliv_next5_di_2","noway_future_improve","mkt_econ_eval_di_1","mkt_econ_eval_di_2")], 2, function(x) {sapply(split(data.frame(df_orig_complete[,"weight"], x), df_orig_complete$group), function(y) weighted.mean(y[,2], w = y$weight))})

Tbl3_rep <- as.data.frame(t(round(Tbl3_rep, 3)))

Tbl3_rep[nrow(Tbl3_rep)+1,] <- cases

rownames(Tbl3_rep)[8] <- "N"

Tbl3_rep[1,] <- round(as.numeric(Tbl3_rep[1,])*100,0)
Tbl3_rep[2,] <- round(as.numeric(Tbl3_rep[2,])*100,0)
Tbl3_rep[3,] <- round(as.numeric(Tbl3_rep[3,])*100,0)
Tbl3_rep[4,] <- round(as.numeric(Tbl3_rep[4,])*100,0)
Tbl3_rep[5,] <- round(as.numeric(Tbl3_rep[5,])*100,0)
Tbl3_rep[6,] <- round(as.numeric(Tbl3_rep[6,])*100,0)
Tbl3_rep[7,] <- round(as.numeric(Tbl3_rep[7,])*100,0)
Tbl3_rep[8,] <- round(as.numeric(Tbl3_rep[8,]),0)

kable_styling(kable(Tbl3_rep, caption = "Table 3R. Replicated Weights. Age-based inequality in economic experience in percentages", col.names = c("Wave 1, Age 18-29", "Wave 1, Age 30-44", "Wave 1, Age 45-59", "Wave 1, 60+", "Wave 2, Age 18-29", "Wave 2, Age 30-44", "Wave 2, Age 45-59", "Wave 2, 60+")))
```

Table 3R. Replicated Weights. Age-based inequality in economic experience in percentages

	Wave 1, Age 18-29	Wave 1, Age 30-44	Wave 1, Age 45-59	Wave 1, 60+	Wave 2, Age 18-29	Wave 2, Age 30-44	Wave 2, Age 45-59	Wave 2, 60+
stdliv_past5_di_1	57	66	75	81	21	43	52	58
stdliv_past5_di_2	20	16	9	5	68	31	25	11

stdliv_next5_di_1	19	25	33	32	11	29	42	38
stdliv_next5_di_2	37	29	22	16	71	32	24	13
noway_future_improve	10	16	35	74	12	16	29	81
mkt_econ_eval_di_1	24	20	17	15	65	17	18	23
mkt_econ_eval_di_2	47	54	57	54	22	57	55	51
N	3827	7325	5766	5226	2086	3596	3876	3946

Here I can reproduce Table 3 almost exactly, but only when I **do not use weights**, therefore, it is unclear what it means in the text when they claim that the descriptives are weighted.

```
Tbl3_rep_n <- apply(df_orig_complete[,c("stdliv_past5_di_1","stdliv_past5_di_2","stdliv_next5_
di_1","stdliv_next5_di_2","noway_future_improve","mkt_econ_eval_di_1","mkt_econ_eval_di_2")],
2, function(x) {sapply(split(data.frame(df_orig_complete[,"weight"], x), df_orig_complete$grou
p), function(y) mean(y[,2], w = y$weight))})

Tbl3_rep_n <- as.data.frame(t(round(Tbl3_rep_n, 3)))

Tbl3_rep_n[nrow(Tbl3_rep_n)+1,] <- cases

rownames(Tbl3_rep_n)[7] <- "N"

Tbl3_rep_n[1,] <- round(as.numeric(Tbl3_rep_n[1,])*100,0)
Tbl3_rep_n[2,] <- round(as.numeric(Tbl3_rep_n[2,])*100,0)
Tbl3_rep_n[3,] <- round(as.numeric(Tbl3_rep_n[3,])*100,0)
Tbl3_rep_n[4,] <- round(as.numeric(Tbl3_rep_n[4,])*100,0)
Tbl3_rep_n[5,] <- round(as.numeric(Tbl3_rep_n[5,])*100,0)
Tbl3_rep_n[6,] <- round(as.numeric(Tbl3_rep_n[6,])*100,0)
Tbl3_rep_n[7,] <- round(as.numeric(Tbl3_rep_n[7,])*100,0)
Tbl3_rep_n[8,] <- round(as.numeric(Tbl3_rep_n[8,]),0)

# create csv for easy importing in final report
write.csv(Tbl3_rep_n, here::here("results", "Tbl3_rep_n.csv"), row.names = F)

kable_styling(kable(Tbl3_rep_n, caption = "Table 3R. Replicated No Weights. Age-based inequali
ty in economic experience in percentages", col.names = c("Wave 1, Age 18-29", "Wave 1, Age 30-
44", "Wave 1, Age 45-59", "Wave 1, 60+", "Wave 2, Age 18-29", "Wave 2, Age 30-44", "Wave 2, Ag
e 45-59", "Wave 2, 60+")))
```

Table 3R. Replicated No Weights. Age-based inequality in economic experience in percentages

	Wave 1, Age 18-29	Wave 1, Age 30-44	Wave 1, Age 45-59	Wave 1, 60+	Wave 2, Age 18-29	Wave 2, Age 30-44	Wave 2, Age 45-59	Wave 2, 60+
stdliv_past5_di_1	57	66	75	81	19	24	34	39

stdliv_past5_di_2	20	16	9	5	48	42	30	20
stdliv_next5_di_1	19	25	33	32	7	11	18	24
stdliv_next5_di_2	37	29	22	16	52	43	31	19
noway_future_improve	10	16	35	74	4	8	24	67
mkt_econ_eval_di_1	24	20	17	15	38	35	30	25
N	47	54	57	54	33	36	42	43
8	3827	7325	5766	5226	2086	3596	3876	3946

Table 3O. Original

```
colnames(Tbl3_orig) <- c("Variable", "Wave 1, Age 18-29", "Wave 1, Age 30-44", "Wave 1, Age 45-59", "Wave 1, 60+", "Wave 2, Age 18-29", "Wave 2, Age 30-44", "Wave 2, Age 45-59", "Wave 2, 60+")

Tbl3_orig$Variable <- c(NA, "stdliv_past5_fallen","stdliv_past5_risen", NA, "stdliv_next5_fallen","stdliv_next5_risen","noway_future_improve",NA, "mkt_econ_eval_pos","mkt_econ_eval_neg", "N")

Tbl3_orig <- subset(Tbl3_orig, !is.na(Variable))

kable_styling(kable(Tbl3_orig, caption = "Table 3O. Original. Age-based inequality in economic experience in percentages"))
```

Table 3O. Original. Age-based inequality in economic experience in percentages

Variable	Wave 1, Age 18-29	Wave 1, Age 30-44	Wave 1, Age 45-59	Wave 1, 60+	Wave 2, Age 18-29	Wave 2, Age 30-44	Wave 2, Age 45-59	Wave 2, 60+
stdliv_past5_fallen	57	66	75	80	19	24	34	40
stdliv_past5_risen	19	15	9	5	49	42	31	19
stdliv_next5_fallen	19	26	34	34	7	12	19	25
stdliv_next5_risen	37	30	23	17	53	43	31	19
noway_future_improve	10	15	35	74	6	10	31	79
mkt_econ_eval_pos	24	20	17	17	42	35	29	24
mkt_econ_eval_neg	45	53	56	52	31	36	42	43
N	3800	6897	5542	5019	2107	3667	3945	4045

Table A.1

Just to check the sample

```
TblA1_repa <- df_orig_complete %>%
  subset(year == 1993) %>%
  group_by(cntry) %>%
  count(cntry) %>%
  ungroup()

TblA1_repb <- df_orig_complete %>%
  subset(year == 1994) %>%
  group_by(cntry) %>%
  count(cntry) %>%
  ungroup()

TblA1_repc <- df_orig_complete %>%
  subset(year == 1996) %>%
  group_by(cntry) %>%
  count(cntry) %>%
  ungroup()

TblA1_repd <- df_orig_complete %>%
  subset(year == 2007) %>%
  group_by(cntry) %>%
  count(cntry) %>%
  ungroup()

TblA1_rep <- left_join(TblA1_repd, TblA1_repa, by = "cntry")
TblA1_rep <- left_join(TblA1_rep, TblA1_repb, by = "cntry")
TblA1_rep <- left_join(TblA1_rep, TblA1_repc, by = "cntry")

TblA1_rep[,1:5] <- TblA1_rep[,c(1,3,4,5,2)]

colnames(TblA1_rep) <- c("country", "1993 N","1994 N","1996 N","2007 N")

TblA1r <- kable_styling(kable(TblA1_rep))

save_kable(TblA1r, file = "results/TblA1_rep.htm")

webshot("results/TblA1_rep.htm", file = "results/TblA1_rep.png")
```

country	1993 N	1994 N	1996 N	2007 N
belarus	1026	NA	NA	918
bulgaria	1671	NA	NA	878
czech	NA	1278	NA	835
estonia	1874	NA	NA	918
hungary	1226	NA	NA	853
latvia	NA	NA	1843	912
lithuania	1794	NA	NA	770
moldova	NA	NA	1536	944
poland	1410	NA	NA	1237
romania	1481	NA	NA	1119
russia	1597	NA	1827	1864
slovakia	NA	1307	NA	915
ukraine	2274	NA	NA	1341

```
kable_styling(kable(TblA1_rep, caption = "Table 1.A Replicated"))
```

Table 1.A Replicated

country	1993 N	1994 N	1996 N	2007 N
belarus	1026	NA	NA	918
bulgaria	1671	NA	NA	878
czech	NA	1278	NA	835
estonia	1874	NA	NA	918
hungary	1226	NA	NA	853
latvia	NA	NA	1843	912
lithuania	1794	NA	NA	770
moldova	NA	NA	1536	944
poland	1410	NA	NA	1237

romania	1481	NA	NA	1119
russia	1597	NA	1827	1864
slovakia	NA	1307	NA	915
ukraine	2274	NA	NA	1341

```
# remove some unused variables to save filesize
df_orig_complete <- df_orig_complete %>%
  dplyr::select(stdliv_next5, stdliv_next5_di, stdliv_past5, stdliv_past5_di, cat_age, wave, f
emale, education, EGP6, income, pensions, unemployed, car_owner, cntry)

write.csv(df_orig_complete, here::here("data/df_orig_complete.csv"))
```

Tech 2. Pre-Analysis 5% Sample.

This is the analysis carried out in the 'Data Finder' phase of the SCORE project used for the pre-registration.

```
df_orig_complete <- read_csv(file = here::here("data/df_orig_complete.csv"))
```

Here I select out 5% of the sample at random and run their models as part of the pre-registration.

```
set.seed(90210)

df_orig_complete <- df_orig_complete %>%
  mutate(rand = runif(NROW(df_orig_complete$y1)),
         p5 = ifelse(rand > 0.05, NA, 1))

df_orig_complete_5 <- df_orig_complete[!is.na(df_orig_complete["p5"]),]
```

```
Tbl5_m1 <- polr(factor(stdliv_next5) ~ factor(cat_age) + factor(wave) + factor(cat_age)*factor
(wave), data = df_orig_complete_5, Hess = T)

Tbl5_m2 <- polr(factor(stdliv_next5) ~ factor(cat_age) + factor(wave) + factor(cat_age)*factor
(wave) + female + factor(education) + factor(EGP6) + factor(income), data = df_orig_complete_5
, Hess = T)

Tbl5_m3 <- polr(factor(stdliv_next5) ~ factor(cat_age) + factor(wave) + factor(cat_age)*factor
(wave) + female + factor(education) + factor(EGP6) + factor(income) + pensions + unemployed +
car_owner + factor(cntry), data = df_orig_complete_5, Hess = T)

tab_model(Tbl5_m1, Tbl5_m2, Tbl5_m3, p.style = "stars", show.ci = F, show.loglik = T,  pred.la
bels = c("k1","k2","k3","k4","Age 30-44","Age 44-59","Age >60","Year 2007","Age 30-44*year'07"
, "Age 44-59*year'07", "Age >60*year'07", "Female", "Educ mid", "Educ high" ,"EGP: routine non
-man", "EGP: Self", "EGP: Skilled", "EGP: Unskilled", "EGP: Farmers", "EGP: Never had a job",
"Income mid", "Income high","Income missing", "Pensions and benefits","Unemployed", "Car","bul
garia", "czech", "estonia", "hungary","latvia", "lithuania", "moldova","poland", "romania", "r
ussia","slovakia", "ukraine"), file = "results/Tbl5_rep.htm")
```

	factor(stdliv next 5)	factor(stdliv next 5)	factor(stdliv next 5)
Predictors	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*
k1	0.08 ***	0.09 ***	0.10 ***
k2	0.20 ***	0.22 ***	0.26 ***
k3	1.72 ***	1.93 **	2.44 **
k4	27.71 ***	32.20 ***	43.24 ***
Age 30-44	0.76	0.74	0.76
Age 44-59	0.46 ***	0.46 ***	0.46 ***
	***	***	***

Tech 2. Pre-Analysis 5% Sample.

Age >60	0.34	0.40	0.45
Year 2007	2.45 ***	2.72 ***	2.91 ***
Age 30-44*year'07	0.63	0.57	0.54 *
Age 44-59*year'07	0.65	0.58	0.57
Age >60*year'07	0.61	0.53 *	0.51 *
Female		0.86	0.89
Educ mid		1.12	1.07
Educ high		1.19	1.23
EGP: routine non-man		0.88	0.94
EGP: Self		1.39	1.41
EGP: Skilled		0.95	0.96
EGP: Unskilled		0.79	0.84
EGP: Farmers		0.98	1.08
EGP: Never had a job		0.54 *	0.67
Income mid		1.10	1.10
Income high		1.52 ***	1.50 ***
Income missing		0.76	0.74
Pensions and benefits			0.93
Unemployed			1.02
Car			1.36 **
bulgaria			0.64
czech			1.30
estonia			1.46
hungary			0.57 *
latvia			2.29 **
lithuania			1.35
moldova			0.56 *
poland			0.94
romania			1.10

russia			1.01
slovakia			0.66
ukraine			0.88
Observations	1774	1774	1774
R^2 Nagelkerke	0.071	0.095	0.141
log-Likelihood	-2266.751	-2245.657	-2202.994

- *p<0.05 ** p<0.01 *** p<0.001*

Output Table

```
#not used in the end

#webshot("results/Tbl5_rep.htm", file = "results/Tbl5_rep.png")

#knitr::include_graphics(here::here("results/Tbl5_rep.png"))
```

Colophon

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] knitr_1.38       webshot_0.5.2    sjPlot_2.8.10    MASS_7.3-55
##  [5] kableExtra_1.3.4 forcats_0.5.1    stringr_1.4.0    dplyr_1.0.8
##  [9] purrr_0.3.4      readr_2.1.2      tidyr_1.2.0      tibble_3.1.6
## [13] ggplot2_3.3.5    tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-155       fs_1.5.2           bit64_4.0.5        lubridate_1.8.0
##  [5] insight_0.17.0     httr_1.4.2         rprojroot_2.0.2    tools_4.1.3
##  [9] backports_1.4.1    bslib_0.3.1        utf8_1.2.2         R6_2.5.1
## [13] sjlabelled_1.1.8   DBI_1.1.2          colorspace_2.0-3   withr_2.5.0
```

```
## [17] tidyselect_1.1.2   emmeans_1.7.3      bit_4.0.4          compiler_4.1.3
## [21] performance_0.9.0  cli_3.2.0          rvest_1.0.2        pacman_0.5.1
## [25] xml2_1.3.3         bayestestR_0.11.5  sass_0.4.1         scales_1.1.1
## [29] mvtnorm_1.1-3      systemfonts_1.0.4  digest_0.6.29      minqa_1.2.4
## [33] rmarkdown_2.13     svglite_2.1.0      pkgconfig_2.0.3    htmltools_0.5.2
## [37] lme4_1.1-28        dbplyr_2.1.1       fastmap_1.1.0      rlang_1.0.2
## [41] readxl_1.4.0       rstudioapi_0.13    jquerylib_0.1.4    generics_0.1.2
## [45] jsonlite_1.8.0     vroom_1.5.7        magrittr_2.0.2     parameters_0.17.0
## [49] Matrix_1.4-0       Rcpp_1.0.8.3       munsell_0.5.0      fansi_1.0.3
## [53] lifecycle_1.0.1    stringi_1.7.6      yaml_2.3.5         snakecase_0.11.0
## [57] grid_4.1.3         parallel_4.1.3     sjmisc_2.8.9       crayon_1.5.1
## [61] lattice_0.20-45    ggeffects_1.1.1    haven_2.4.3        splines_4.1.3
## [65] sjstats_0.18.1     hms_1.1.1          pillar_1.7.0       boot_1.3-28
## [69] estimability_1.3   effectsize_0.6.0.1 reprex_2.0.1       glue_1.6.2
## [73] evaluate_0.15      modelr_0.1.8       vctrs_0.3.8        nloptr_2.0.0
## [77] tzdb_0.3.0         cellranger_1.1.0   gtable_0.3.0       assertthat_0.2.1
## [81] datawizard_0.4.0   xfun_0.30          xtable_1.8-4       broom_0.7.12
## [85] viridisLite_0.4.0  ellipsis_0.3.2     here_1.0.1
```

03 Main Analysis

Nate Breznau

2022-04-15

```
df_orig_complete <- read_csv(file = here::here("data", "df_orig_complete.csv"))
```

```
Tbl5_m1 <- polr(factor(stdliv_next5) ~ factor(cat_age) + factor(wave) + factor(cat_age)*factor
(wave), data = df_orig_complete, Hess = T)

Tbl5_m2 <- polr(factor(stdliv_next5) ~ factor(cat_age) + factor(wave) + factor(cat_age)*factor
(wave) + female + factor(education) + factor(EGP6) + factor(income), data = df_orig_complete,
Hess = T)

Tbl5_m3 <- polr(factor(stdliv_next5) ~ factor(cat_age) + factor(wave) + factor(cat_age)*factor
(wave) + female + factor(education) + factor(EGP6) + factor(income) + pensions + unemployed +
car_owner + factor(cntry), data = df_orig_complete, Hess = T)

tab_model(Tbl5_m1, Tbl5_m2, Tbl5_m3, transform = NULL, p.style = "stars", show.ci = F, show.lo
glik = T,  pred.labels = c("k1","k2","k3","k4","Age 30-44","Age 44-59","Age >60","Year 2007","
Age 30-44*year'07", "Age 44-59*year'07", "Age >60*year'07", "Female", "Educ mid", "Educ high"
,"EGP: routine non-man", "EGP: Self", "EGP: Skilled", "EGP: Unskilled", "EGP: Farmers", "EGP:
Never had a job", "Income mid", "Income high","Income missing", "Pensions and benefits","Unemp
loyed", "Car","bulgaria", "czech", "estonia", "hungary","latvia", "lithuania", "moldova","pola
nd", "romania", "russia","slovakia", "ukraine"), file = here::here("results", "Tbl5_rep_f.doc"
))
```

	factor(stdliv next 5)	factor(stdliv next 5)	factor(stdliv next 5)
Predictors	*Log-Odds*	*Log-Odds*	*Log-Odds*
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 ***
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 ***
bulgaria			0.02
czech			0.33 ***
estonia			0.66 ***
hungary			-0.24 ***
latvia			0.72 ***
lithuania			0.58 ***
moldova			-0.27 ***
poland			0.27 ***
romania			0.40 ***
russia			0.15 **
slovakia			0.00
ukraine			-0.12 *
Observations	35648	35648	35648
R^2 Nagelkerke	0.068	0.086	0.120

log-Likelihood	-45584.674	-45258.680	-44644.165

- *p<0.05 ** p<0.01 *** p<0.001*

```
tab_model(Tbl5_m1, Tbl5_m2, Tbl5_m3, transform = NULL, p.style = "stars", show.ci = F, show.loglik = T,  pred.labels = c("k1","k2","k3","k4","Age 30-44","Age 44-59","Age >60","Year 2007","Age 30-44*year'07", "Age 44-59*year'07", "Age >60*year'07", "Female", "Educ mid", "Educ high","EGP: routine non-man", "EGP: Self", "EGP: Skilled", "EGP: Unskilled", "EGP: Farmers", "EGP: Never had a job", "Income mid", "Income high","Income missing", "Pensions and benefits","Unemployed", "Car","bulgaria", "czech", "estonia", "hungary","latvia", "lithuania", "moldova","poland", "romania", "russia","slovakia", "ukraine"), file = here::here("results", "Tbl5_rep_f.htm"))
```

	factor(stdliv next 5)	**factor(stdliv next 5)**	**factor(stdliv next 5)**
Predictors	*Log-Odds*	*Log-Odds*	*Log-Odds*
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 ***
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 ***
bulgaria			0.02
czech			0.33 ***
estonia			0.66 ***
hungary			-0.24 ***
latvia			0.72 ***
lithuania			0.58 ***
moldova			-0.27 ***
poland			0.27 ***
romania			0.40 ***
russia			0.15 **
slovakia			0.00
ukraine			-0.12 *
Observations	35648	35648	35648
R^2 Nagelkerke	0.068	0.086	0.120
log-Likelihood	-45584.674	-45258.680	-44644.165

- *p<0.05 ** p<0.01 *** p<0.001*

```
#knitr::include_graphics(here::here("results", "Tbl5_rep_f.htm"))

#webshot(here::here("results", "Tbl5_rep_f.htm"), file = here::here("results", "Tbl5_rep_f.png
"))

#knitr::include_graphics(here::here("results", "Tbl5_rep_f.png"))
```

Marginal Estimates

The coefficients in the table are different from the original results, but they are also different from the Stata results (see 03_Main_Analysis_Stata.do). Clearly the ordred probit models use slightly different estimation techniques. But, the

quesiton is: are the predicted values the same?

Calculate Margins

```
newdat <- data.frame(
  cat_age = rep(1:4, 2),
  wave = c(rep(1993, 4), rep(2007, 4)),
  female = rep(0, 8),
  education = rep(1,8),
  EGP6 = rep(1,8),
  income = rep(1,8),
  pensions = rep(mean(df_orig_complete$pensions, na.rm = T),8),
  unemployed = rep(mean(df_orig_complete$unemployed, na.rm = T),8),
  car_owner = rep(mean(df_orig_complete$car_owner, na.rm = T),8),
  cntry = "belarus") # ref cat

newdat <- cbind(newdat, predict(Tbl5_m3, newdat, type = "probs"))
```

Import Stata Results

```
Tbl5_m3_Stata <- read_csv(here::here("results","Tbl5_m3.csv"))
```

```
## Rows: 40 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (5): outcome, age_cat, wave, margin, se
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Combine into Margins Comparison Table

```
Tbl5_marg_comp <- newdat %>%
  dplyr::select(wave, cat_age, `1`, `2`, `3`, `4`, `5`)

colnames(Tbl5_marg_comp) <- c("wave", "age_cat", "fall_alot", "fall", "stay_same", "rise", "ri
se_alot")

# add Stata results
Tbl5_marg_comp[9:16,] <- NA
Tbl5_marg_comp$Software <- c(rep("R polr",8), rep("Stata oprobit", 8))
Tbl5_marg_comp[9:16,1:2] <- Tbl5_marg_comp[1:8,1:2]

#sort Stata data
Tbl5_m3_Stata <- Tbl5_m3_Stata[order( Tbl5_m3_Stata[,"outcome"], Tbl5_m3_Stata[,"wave"] ),]

Tbl5_marg_comp$fall_alot[9:16] <- Tbl5_m3_Stata$margin[1:8]
Tbl5_marg_comp$fall[9:16] <- Tbl5_m3_Stata$margin[9:16]
Tbl5_marg_comp$stay_same[9:16] <- Tbl5_m3_Stata$margin[17:24]
Tbl5_marg_comp$rise[9:16] <- Tbl5_m3_Stata$margin[25:32]
```

```
Tbl5_marg_comp$rise_alot[9:16] <- Tbl5_m3_Stata$margin[33:40]
```

Comparison

```
Tbl5_marg_comp %>%
  subset(age_cat == 1 | age_cat == 4) %>%
  ggplot(aes(y = (fall_alot+fall), x = wave, color = interaction(Software,age_cat))) +
  geom_point() +
  geom_line() +
  labs(y = "Predicted Likelihood of Percieved Standard\nof Living Falling in the Next 5 Years"
, x = "Survey Wave") +
  scale_color_manual(name = "Age group &\nSoftware used",
                     values = c("#481567FF",
                                "#238A8DFF",
                                "#453781FF",
                                "#29AF7FFF"),
                     labels = c("Age 19-29\nusing R polr", "Age 19-29\nusing Stata oprobit",
 "Age 60+\nusing R polr", "Age 60+\nusing Stata oprobit")) +
  scale_x_discrete(limits = c(1993,2007)) +
  theme_classic()
```

```
## Warning: Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```

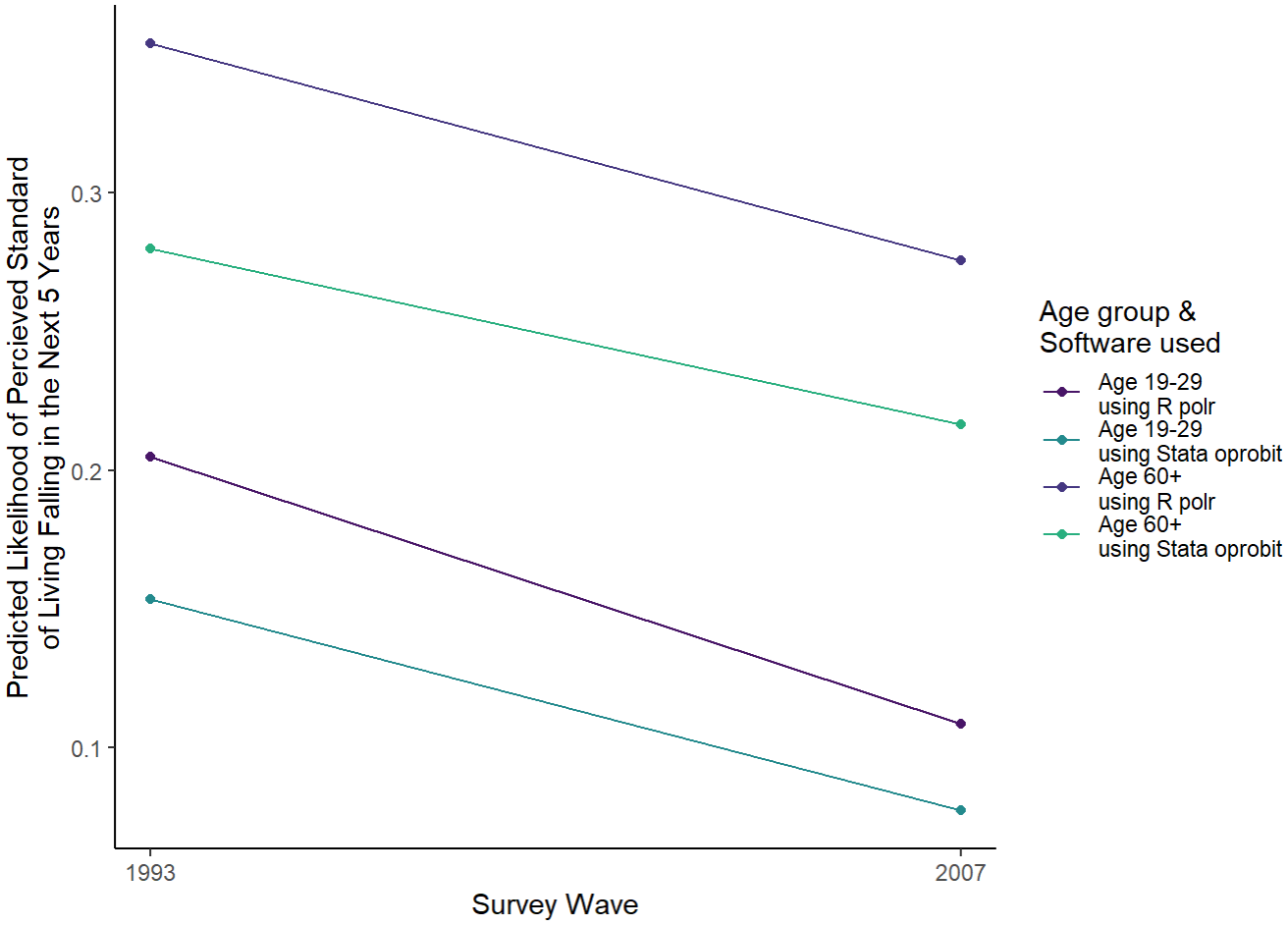

Table Comparing Stata and R Margins

```
kable_styling(kable(Tbl5_marg_comp))
```

wave	age_cat	fall_alot	fall	stay_same	rise	rise_alot	Software
1993	1	0.0927751	0.1120754	0.5111955	0.2567204	0.0272337	R polr
1993	2	0.1396044	0.1505535	0.5098853	0.1826181	0.0173387	R polr
1993	3	0.1888147	0.1808238	0.4819878	0.1362233	0.0121504	R polr
1993	4	0.1784101	0.1752032	0.4890246	0.1443496	0.0130125	R polr
2007	1	0.0460742	0.0624050	0.4351127	0.4004501	0.0559580	R polr
2007	2	0.0722499	0.0917628	0.4935636	0.3069644	0.0354592	R polr
2007	3	0.1088939	0.1264952	0.5154429	0.2262760	0.0228920	R polr
2007	4	0.1311603	0.1443632	0.5127259	0.1931386	0.0186120	R polr
1993	1	0.0626676	0.0909308	0.4790867	0.3246091	0.0427057	Stata oprobit
1993	2	0.1054630	0.1243734	0.5027682	0.2447451	0.0226503	Stata oprobit
1993	3	0.1464445	0.1480962	0.4993305	0.1922632	0.0138655	Stata oprobit
1993	4	0.1366441	0.1430050	0.5015386	0.2033232	0.0154891	Stata oprobit
2007	1	0.0265037	0.0508320	0.3974964	0.4314184	0.0937495	Stata oprobit
2007	2	0.0479969	0.0764672	0.4575851	0.3618393	0.0561114	Stata oprobit
2007	3	0.0808522	0.1064922	0.4943244	0.2864267	0.0319044	Stata oprobit
2007	4	0.0974332	0.1188876	0.5011741	0.2573054	0.0251997	Stata oprobit

Colophon

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
```

```
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] knitr_1.38       webshot_0.5.2    sjPlot_2.8.10    MASS_7.3-55
##  [5] kableExtra_1.3.4 forcats_0.5.1    stringr_1.4.0    dplyr_1.0.8
##  [9] purrr_0.3.4      readr_2.1.2      tidyr_1.2.0      tibble_3.1.6
## [13] ggplot2_3.3.5    tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-155       fs_1.5.2           bit64_4.0.5        lubridate_1.8.0
##  [5] insight_0.17.0     httr_1.4.2         rprojroot_2.0.2    tools_4.1.3
##  [9] backports_1.4.1    bslib_0.3.1        utf8_1.2.2         R6_2.5.1
## [13] sjlabelled_1.1.8   DBI_1.1.2          colorspace_2.0-3   withr_2.5.0
## [17] tidyselect_1.1.2   emmeans_1.7.3      bit_4.0.4          compiler_4.1.3
## [21] performance_0.9.0  cli_3.2.0          rvest_1.0.2        pacman_0.5.1
## [25] xml2_1.3.3         labeling_0.4.2     bayestestR_0.11.5  sass_0.4.1
## [29] scales_1.1.1       mvtnorm_1.1-3      systemfonts_1.0.4  digest_0.6.29
## [33] minqa_1.2.4        rmarkdown_2.13     svglite_2.1.0      pkgconfig_2.0.3
## [37] htmltools_0.5.2    lme4_1.1-28        highr_0.9          dbplyr_2.1.1
## [41] fastmap_1.1.0      rlang_1.0.2        readxl_1.4.0       rstudioapi_0.13
## [45] farver_2.1.0       jquerylib_0.1.4    generics_0.1.2     jsonlite_1.8.0
## [49] vroom_1.5.7        magrittr_2.0.2     parameters_0.17.0  Matrix_1.4-0
## [53] Rcpp_1.0.8.3       munsell_0.5.0      fansi_1.0.3        lifecycle_1.0.1
## [57] stringi_1.7.6      yaml_2.3.5         snakecase_0.11.0   grid_4.1.3
## [61] parallel_4.1.3     sjmisc_2.8.9       crayon_1.5.1       lattice_0.20-45
## [65] ggeffects_1.1.1    haven_2.4.3        splines_4.1.3      sjstats_0.18.1
## [69] hms_1.1.1          pillar_1.7.0       boot_1.3-28        estimability_1.3
## [73] effectsize_0.6.0.1 reprex_2.0.1       glue_1.6.2         evaluate_0.15
## [77] modelr_0.1.8       vctrs_0.3.8        nloptr_2.0.0       tzdb_0.3.0
## [81] cellranger_1.1.0   gtable_0.3.0       assertthat_0.2.1   datawizard_0.4.0
## [85] xfun_0.30          xtable_1.8-4       broom_0.7.12       viridisLite_0.4.0
## [89] ellipsis_0.3.2     here_1.0.1
```