

Generate your own SHARE dataset

Stefan Gruber (MEA)

SHARE User Workshop
Stockholm 25/02/2016

Generate your own panel dataset

1. Step: Define your research question and the research design
2. Step: Which modules and variables do you need to answer your research question?
3. Step: Generate your panel dataset and prepare the data
4. Step: Analysis

Step 1: Define your research question and the research design

Step 1: Define research question and design

Research question:

Does retirement influence the individual
quality of life?

Step 1: Define research question and design

The model:

- ▶ **Dependent variable:** quality of life (CASP)
- ▶ **Independent variable:** retirement
- ▶ **Possible controls:** age, health status, marital status, educational level, financial situation

Step 2: Which modules and variables do you need to answer your research question?

Step 2: Which modules and variables do you need to answer your research question?

Most important:
Become acquainted with the dataset!!

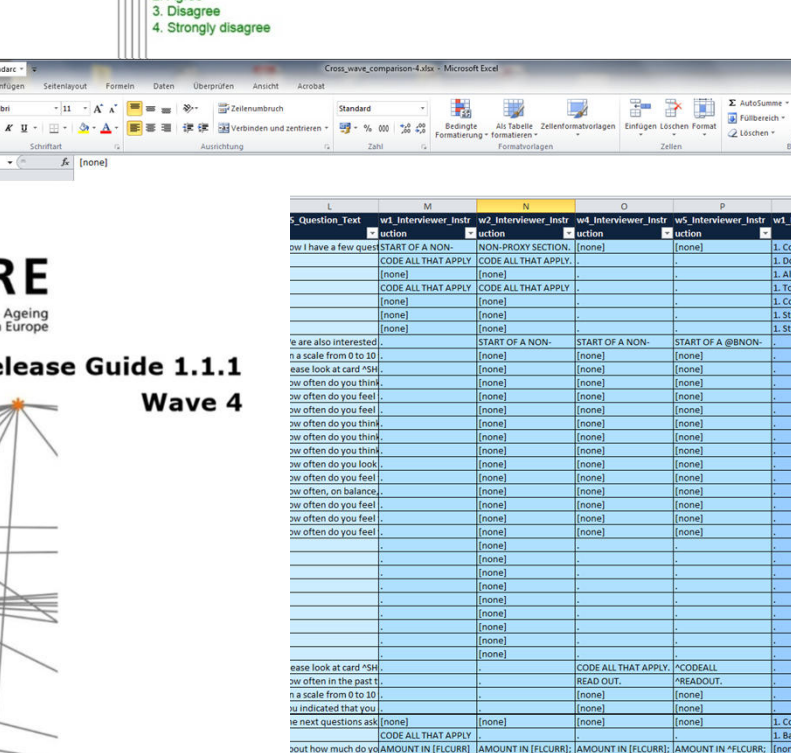
Step 2: Which modules and variables do you need to answer your research question?

1. Study the questionnaires

2. Cross-Wave-Comparison-Tool

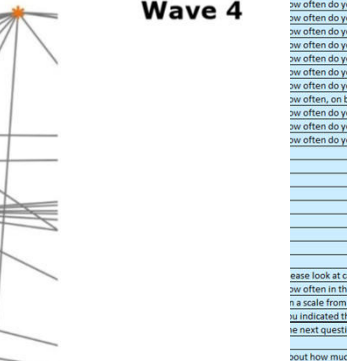
3. Release Guides

EP029_LitFreeWork
I have very little freedom to decide how I do my work. (Would you say you strongly agree, agree, disagree or strongly disagree?)
IWER: Show card 22
 1. Strongly agree
 2. Agree
 3. Disagree
 4. Strongly disagree



SHARE
 Survey of Health, Ageing
 and Retirement in Europe

Release Guide 1.1.1
Wave 4

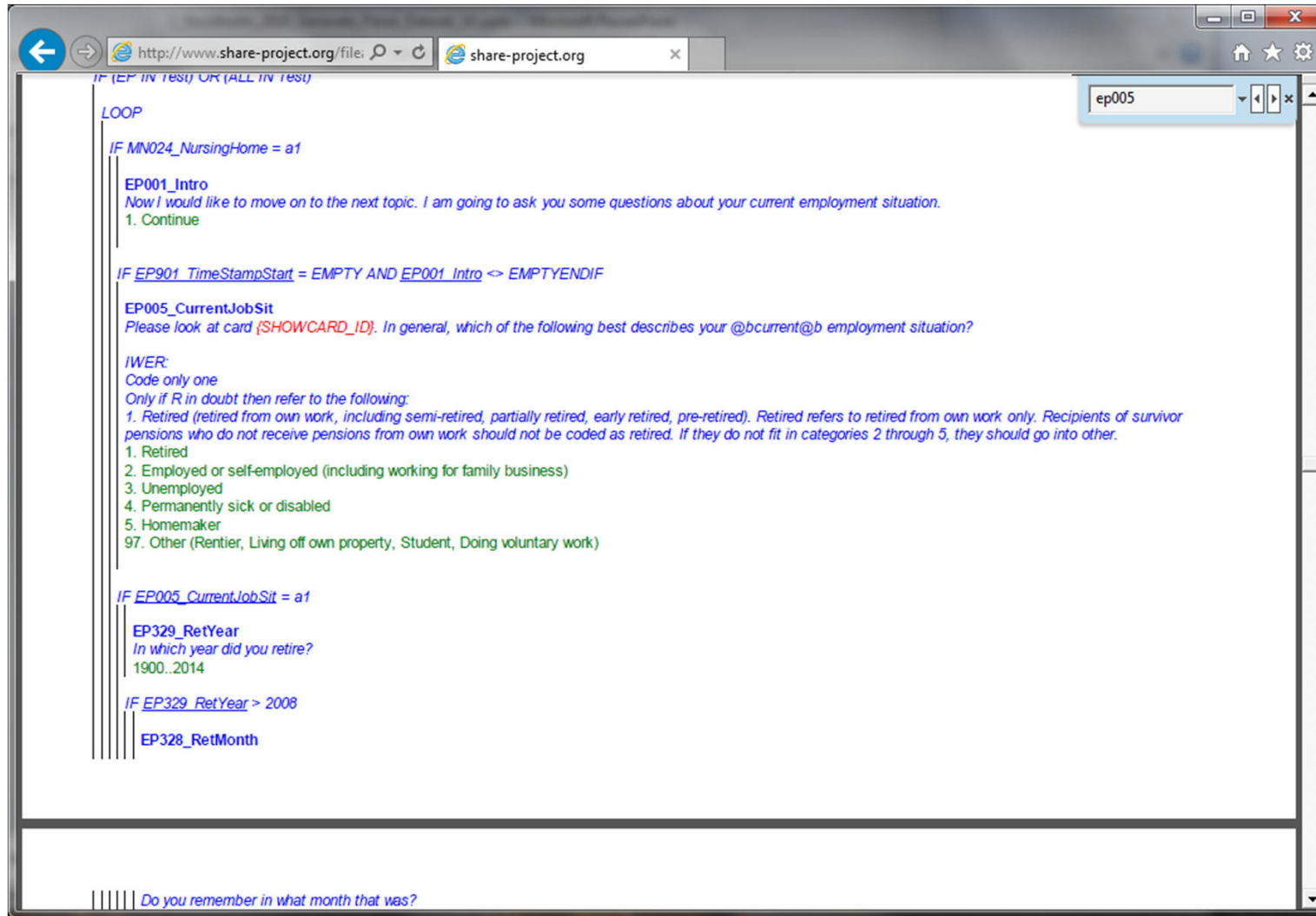


Step 2: Which modules and variables do you need to answer your research question?

Quality of life in SHARE: **CASP-12**

- ▶ The score measures quality of life and well-being
- ▶ It is based on four subscales: *control*, *autonomy*, *self-realization* and *pleasure*.
- ▶ The score is the sum of these four subscales and ranges from 12 to 48.
- ▶ The items were asked in the drop off questionnaire in wave 1 and then moved into the CAPI from wave 2 onwards.
- ▶ For drop off questionnaires not being filled in by all respondents there is a higher fraction of missing observations in wave 1.

Step 2: Which modules and variables do you need to answer your research question?



IF (EP IN Test) OR (ALL IN Test)

LOOP

IF MN024_NursingHome = a1

EP001_Intro
Now I would like to move on to the next topic. I am going to ask you some questions about your current employment situation.

1. Continue

IF EP901_TimeStampStart = EMPTY AND EP001_Intro <> EMPTYENDIF

EP005_CurrentJobSit
Please look at card {SHOWCARD_ID}. In general, which of the following best describes your @bcurrent@b employment situation?

IWER:
Code only one
Only if R in doubt then refer to the following:

1. Retired (retired from own work, including semi-retired, partially retired, early retired, pre-retired). Retired refers to retired from own work only. Recipients of survivor pensions who do not receive pensions from own work should not be coded as retired. If they do not fit in categories 2 through 5, they should go into other.
2. Employed or self-employed (including working for family business)
3. Unemployed
4. Permanently sick or disabled
5. Homemaker
97. Other (Rentier, Living off own property, Student, Doing voluntary work)

IF EP005_CurrentJobSit = a1

EP329_RetYear
In which year did you retire?
1900..2014

IF EP329_RetYear > 2008

EP328_RetMonth

Do you remember in what month that was?

Step 2: Which modules and variables do you need to answer your research question?

Which modules do we need:

- ▶ Coverscreen (cv_r) → age
- ▶ Dropoff module of wave 1 → CASP
- ▶ Activities module (ac) of waves 2, 4 & 5 → CASP
- ▶ Demographics (dn) → marital status
- ▶ Consumption (co) → financial situation (able to make ends meet)
- ▶ Employment (ep) → retirement
- ▶ Gv_isced → educational level
- ▶ Gv_health → health status

Step 3: Generate your panel dataset and prepare the data

Step 3: Generate panel dataset and prepare the data

1. Extract & Recode Variables from the wave-specific modules
2. Merge modules per wave → *Attention: cv_r also contains non-responding household members!*
3. Copying information collected only by one hh-member
→ *CO module answered by hh-respondent only*
4. Append waves to panel long format
5. Copying information collected only once in baseline interview to later waves (ISCED)

```
asySHARE_rel2-0-0.do x Panel_dataset_exercise.do x Generate_Panel.do x Untitled.do x

.6
.7 *----[6. Extract & Recode Variables from CO ]-----
.8
.9
.10 *>> w1:
.11
.12 use $wave1\sharew1_rel2-6-0_co.dta, clear
.13     tab co007_
.14     gen co007_mod = co007_
.15     replace co007_mod = 0 if co007_ == 3 | co007_ == 4
.16     replace co007_mod = 1 if co007_ == 1 | co007_ == 2
.17     keep mergeid co007_mod
.18     save $temp\sharew1_co_a.dta, replace
.19
.20 *>> w2:
.21
.22 use $wave2\sharew2_rel2-6-0_co.dta, clear
.23     gen co007_mod = co007_
.24     replace co007_mod = 0 if co007_ == 3 | co007_ == 4
.25     replace co007_mod = 1 if co007_ == 1 | co007_ == 2
.26     keep mergeid co007_mod
.27     save $temp\sharew2_co_a.dta, replace
.28
.29 *>> w4:
.30
.31 use $wave4\sharew4_rel1-1-1_co.dta, clear
.32     gen co007_mod = co007_
.33     replace co007_mod = 0 if co007_ == 3 | co007_ == 4
.34     replace co007_mod = 1 if co007_ == 1 | co007_ == 2
.35     keep mergeid co007_mod
.36     save $temp\sharew4_co_a.dta, replace
.37
.38 *>> w5:
.39
.40 use $wave5\sharew5_rel1-0-0_co.dta, clear
.41     gen co007_mod = co007_
.42     replace co007_mod = 0 if co007_ == 3 | co007_ == 4
.43     replace co007_mod = 1 if co007_ == 1 | co007_ == 2
.44     keep mergeid co007_mod
.45     save $temp\sharew5_co_a.dta, replace
.46
```


Do-file Editor - Generate_Panel.do

File Edit View Project Tools

easySHARE_rel2-0-0.do X Panel_dataset_exercise.do X Generate_Panel.do X Untitled.do X

```
478
479
480 *----[10. Merge modules per wave ]-----
481
482 // We use the CV_R modules as master and then merge the other modules.
483
484 *>> w1:
485
486 use $temp\sharew1_cv_r_a.dta, clear
487     assert mergeid!="no int w.1" // dropped non responding hh members above
488     merge 1:1 mergeid using $temp\sharew1_dn_a.dta, assert( 3) nogen
489     merge 1:1 mergeid using $temp\sharew1_co_a.dta, assert( 3) nogen
490     merge 1:1 mergeid using $temp\sharew1_ep_a.dta, assert( 3) nogen
491     merge 1:1 mergeid using $temp\sharew1_gv_health_a.dta, assert( 3) nogen
492     merge 1:1 mergeid using $temp\sharew1_gv_isced_a.dta, assert( 3) nogen
493     merge 1:1 mergeid using $temp\sharew1_dropoff_a.dta, ///
494     gen(merge_w1_dropoff) assert(1 3) // respondents without dropoff
495     drop merge_w1_dropoff
496     save $temp\sharew1_merged_a.dta, replace
497
498 *>> w2:
499
500 use $temp\sharew2_cv_r_a.dta, clear
501     assert mergeid!="no int w.2" // dropped non responding hh members above
502     merge 1:1 mergeid using $temp\sharew2_dn_a.dta, assert( 3) nogen
503     merge 1:1 mergeid using $temp\sharew2_ac_a.dta, assert( 3) nogen
504     merge 1:1 mergeid using $temp\sharew2_co_a.dta, assert( 3) nogen
505     merge 1:1 mergeid using $temp\sharew2_ep_a.dta, assert( 3) nogen
506     merge 1:1 mergeid using $temp\sharew2_gv_health_a.dta, assert( 3) nogen
507     merge 1:1 mergeid using $temp\sharew2_gv_isced_a.dta, assert( 3) nogen
508     save $temp\sharew2_merged_a.dta, replace
509
510 *>> w4:
511
512 use $temp\sharew4_cv_r_a.dta, clear
513     assert mergeid!="no int w.4" // dropped non responding hh members above
514     merge 1:1 mergeid using $temp\sharew4_dn_a.dta, assert(3) nogen
515     merge 1:1 mergeid using $temp\sharew4_ac_a.dta, assert(3) nogen
516     merge 1:1 mergeid using $temp\sharew4_co_a.dta, assert(3) nogen
517     merge 1:1 mergeid using $temp\sharew4_ep_a.dta, assert(3) nogen
518     merge 1:1 mergeid using $temp\sharew4_gv_health_a.dta, assert(3) nogen
519     merge 1:1 mergeid using $temp\sharew4_gv_isced_a.dta, assert(3) nogen
520     save $temp\sharew4_merged_a.dta, replace
521
```



```
535 numlabel, add
536
537
538 *----[11. Copying information collected only by hh-respdent (dumhhr) ]-----
539
540 *>> w1:
541
542 use $temp\sharew1_merged_a.dta, clear
543 *br mergeid hhid1 country dumhhr co007
544 gsort hhid1 -dumhhr // set hh-respondent always on first position
545 replace co007 = co007[_n-1] if co007==. & hhid1 == hhid1[_n-1]
546 mvdecode co007, mv(-1/-2)
547 *br mergeid hhid1 country dumhhr co007
548 save $temp\sharew1_merged_b.dta, replace
549
550 *>> w2:
551
552 use $temp\sharew2_merged_a.dta, clear
553 gsort hhid2 -dumhhr
554 replace co007 = co007[_n-1] if co007==. & hhid2 == hhid2[_n-1]
555 mvdecode co007, mv(-1/-2)
556 save $temp\sharew2_merged_b.dta, replace
557
558 *>> w4:
559
560 use $temp\sharew4_merged_a.dta, clear
561 gsort hhid4 -dumhhr
562 replace co007 = co007[_n-1] if co007==. & hhid4 == hhid4[_n-1]
563 mvdecode co007, mv(-1/-2)
564 save $temp\sharew4_merged_b.dta, replace
565
566
567 *>> w5:
568
569 use $temp\sharew5_merged_a.dta, clear
570 gsort hhid5 -dumhhr
571 replace co007 = co007[_n-1] if co007==. & hhid5 == hhid5[_n-1]
572 mvdecode co007, mv(-1/-2)
573 save $temp\sharew5_merged_b.dta, replace
574
575
```


Do-file Editor - Generate_Panel.do*

File Edit View Project Tools

easySHARE_rel2-0-0.do X Panel_dataset_exercise.do X Generate_Panel.do* X Untitled.do X

```
574
575
576
577 *----[12. Append waves to panel long format & integrate "long" variables]-----
578
579 *>> Append single wave files to one long file:
580
581 use                $temp\sharew5_merged_b.dta, clear // labels
582 append using       $temp\sharew4_merged_b.dta
583 append using       $temp\sharew2_merged_b.dta
584 append using       $temp\sharew1_merged_b.dta
585
586
587 *>> Integrate hhid? into a "long" format hidd variable
588     gen          hhid = ""
589     foreach w in $w {
590         replace hhid = hhid`w' if wave==`w'
591     }
592     lab var  hhid "household identifier in respective wave - see var. wave"
593
594     sort mergeid hhid wave
595     drop hhid?
596
597
598 *>> Copy isced_r from baseline to later waves
599     tab isced_r wave,m
600     sort mergeid wave
601     replace isced_r = isced_r[_n-1] if isced_r ==. & mergeid == mergeid[_n-1]
602     tab isced_r wave,m
603
604
605
606 *----[13. Analysis]-----
607
608     mvdecode casp Retired, mv(-1/-2)
609
610 *>> Cross-sectional analysis with wave 5
611
612
```

Line: 595, Col: 15 CAP NUM OVR

Step 4: Analysis

(exemplary! Do not cite!)

Step 4: Analysis

```
. ta Retired
```

RECODE of ep005_ (Current job situation)	Freq.	Percent	Cum.
0	87,894	46.53	46.53
1	101,014	53.47	100.00
Total	188,908	100.00	

53 percent of the
observations are
retired

```
. sum casp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
casp	168,633	37.36559	6.325392	12	48

Less observations
available for CASP due
to w1 drop-off

Step 4: Analysis

* Cross-sectional analysis:

```
reg casp Retired female age co007_mod sphus_2 sphus_3 sphus_4 ///
sphus_5 isced_r_2 isced_r_3 isced_r_4 isced_r_5 isced_r_6 if wave ==5
```

Source	SS	df	MS	Number of obs	=	21,442
Model	272568.163	13	20966.7818	F(13, 21428)	=	886.82
Residual	506614.269	21,428	23.6426297	Prob > F	=	0.0000
				R-squared	=	0.3498
				Adj R-squared	=	0.3494
Total	779182.433	21,441	36.3407692	Root MSE	=	4.8624

casp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Retired	.2428614	.0707843	3.43	0.001	.1041189 .381604
female	-.166942	.067423	-2.48	0.013	-.2990961 -.0347878
age	-.0034851	.0013703	-2.54	0.011	-.0061711 -.0007992
co007_mod	-3.832087	.0767923	-49.90	0.000	-3.982605 -3.681568
sphus_2	-1.257869	.1239858	-10.15	0.000	-1.50089 -1.014847
sphus_3	-2.579503	.1146536	-22.50	0.000	-2.804233 -2.354774
sphus_4	-5.070172	.1243151	-40.78	0.000	-5.313839 -4.826505
sphus_5	-9.149092	.1580605	-57.88	0.000	-9.458903 -8.839282
isced_r_2	1.822029	.1548813	11.76	0.000	1.51845 2.125608
isced_r_3	2.470873	.1477878	16.72	0.000	2.181198 2.760548
isced_r_4	2.934816	.1384377	21.20	0.000	2.663467 3.206164
isced_r_5	3.275441	.2094251	15.64	0.000	2.864952 3.685929
isced_r_6	3.021708	.1447031	20.88	0.000	2.738079 3.305337
_cons	40.34844	.1919019	210.26	0.000	39.9723 40.72459

Positive significant coefficient: retired people have higher levels of well-being.

Step 4: Analysis

- * To fully exploit SHARE, i.e. estimating a panel model, we first need to `xtset` the data.

- * Unfortunately Stata does not accept numerical person IDs, but there is a simple work-around:

```
bysort mergeid : gen mergeid_n = _n == 1  
replace                mergeid_n = sum(mergeid_n)
```

- * Then we can declare the data to be a panel

```
xtset mergeid_n wave
```

- * Then we can estimate a fixed effects panel model:

```
xtreg casp Retired age co007_mod sphus i.wave, fe
```

Step 4: Analysis

Fixed-effects (within) regression
Group variable: mergeid_n

Number of obs = 165,568
Number of groups = 99,835

R-sq:

within = 0.0464
between = 0.0012
overall = 0.0038

Obs per group:

min = 1
avg = 1.7
max =

corr(u_i, Xb) = -0.5480

F(7,65726) =
Prob > F =

	Coef.	Std. Err.	t	P> t	[95% Conf. In	
casp						
Retired	.5460179	.0553519	9.86	0.000	.4375281	.
age	.2874831	.0400189	7.18	0.000	.2090461	.
co007_mod	-1.387598	.0447612	-31.00	0.000	-1.47533	-1
sphus	-.9763644	.0220599	-44.26	0.000	-1.019602	-
wave						
2	-.5044401	.1101902	-4.58	0.000	-.7204128	-.2884674
4	-1.800124	.2721538	-6.61	0.000	-2.333545	-1.266703
5	-2.221101	.3474725	-6.39	0.000	-2.902147	-1.540054
_cons	23.33125	2.385854	9.78	0.000	18.65498	28.00753
sigma_u	7.6067326					
sigma_e	3.7725325					
rho	.80259209	(fraction of variance due to u_i)				

F test that all u_i=0: F(99834, 65726) = 2.50

Prob > F = 0.0000

The cross-sectional correlation we found still subsists in the panel model: Respondents who retire during the observation period show significantly higher well-being

easySHARE Syntax as example

Do-file Editor - easySHARE_rel2-0-0.do

File Edit View Project Tools

easySHARE_rel2-0-0.do X Untitled.do X

```

46 *---[ Overview of Contents ]-----
47
48 *---[ 0. Stata Version & Settings]-----
49
50 *---[ 1. Define paths and open log file]-----
51
52 *---[ 2. Define w, m, c_* globals]-----
53
54 *---[ 3. Copy Main SHARE Release to easySHARE directory]-----
55
56 *---[ 4. Generate CoupleIDs based on cvid/cvidp]-----
57
58 *---[ 5. Extract & Recode Variables from cv_r]-----
59 *---[ 6. Extract & Recode Variables from DN / ST ]-----
60 *---[ 7. Extract & Recode Variables from AC ]-----
61 *---[ 8. Extract & Recode Variables from BR ]-----
62 *---[ 9. Extract & Recode Variables from CF]-----
63 *---[10. Extract & Recode Variables from CH ]-----
64 *---[11. Extract & Recode Variables from CO ]-----
65 *---[12. Extract & Recode Variables from CS ]-----
66 *---[13. Extract & Recode Variables from EP ]-----
67 *---[14. Extract & Recode Variables from GS ]-----
68 *---[15. Extract & Recode Variables from HC ]-----
69 *---[16. Extract & Recode Variables from MC ]-----
70 *---[17. Extract & Recode Variables from PH ]-----
71 *---[18. Extract & Recode Variables from SN ]-----
72 *---[19. Extract & Recode Variables from Sp ]-----
73 *---[20. Extract & Recode Variables from GV_Health]-----
74 *---[21. Extract & Recode Variables from GV_ISCED ]-----
75 *---[22. Extract & Recode Variables from DROPOFF]-----
76 *---[23. Extract & Recode Variables from imputations]-----
77
78 *---[24. Merge modules per wave ]-----
79
80 *---[25. Other recodes per wave ]-----
81
82 *---[26. Append waves to panel long format & integrate "long" variables]-----
83
84 *---[27. Fix date intv., year/month birth, gender, & partnervars & gen age]-----
85
86 *---[28. Transfer information collected once (in baseline interviews)]-----
87
88 *---[29. Pass on information to next wave that may have changed/not changed]-----
89
90 *---[30. Fix & re-generate variables, labels, etc.]-----
91
92 *---[31. Implement/complete wave/country skip patterns]-----
93
94 *---[32. Integrate DK/RF and implement no information missing code]-----
95
96 *---[33. Keep, add easy missing codes & labels, order, data labels & save]-----
97

```

Line: 268, Col: 61 CAP NUM OVR

Gruber, Stefan; Hunkler, Christian; Stuck, Stephanie (2014):
Generating *easy*SHARE: Guidelines, Structure, Content and
Programming , SHARE Working Paper Series: 17-2014

→ <http://www.share-project.org/publications/share-working-paper-series.html>

Abstract

This paper provides an overview of the construction, structure and content of *easy*SHARE – a simplified dataset that is based on the scientific release of the Survey of Health, Ageing and Retirement in Europe (SHARE). *easy*SHARE contains a subset of ready-to-analyse variables and indices. It is designed for student training on country-comparative and longitudinal analyses as well as for introducing SHARE to researchers who have little experience in quantitative analyses of complex survey data. This paper puts special emphasis on the extraction and generation of the dataset using the statistical software package Stata. We also provide an example of how to extract more variables from the scientific release of SHARE to include them in the *easy*SHARE dataset.

Thank you!

gruber@mea.mpisoc.mpg.de

