FP7 - Design Study

Deliverable 25: Report on the study fieldwork experience, methodological experiments and functionality of newly developed instruments

Please cite this deliverable as:
GGP (2012) Report on the study fieldwork experience, methodological experiments and functionality of newly developed instruments, Deliverable 25 of the EU-DG Research grant entitled '*Design Studies for Research Infrastructures*' funded under the 7th Framework Programme (FP7), GGP 212749. Available at: www.ggp-i.org

GGP 212749

Deliverable 25

Report on the study fieldwork experience, methodological experiments and functionality of newly developed instruments

Gregor Petrič, Katja Lozar Manfreda and Rok Platinovšek
University of Ljubljana

Work Package 13
Survey Instrument Experiments

December 2012

EUROPEAN
COMMISSION

European
Research Area

Contents

1. Introduction

The main goal of this report is to present the results of statistical analyses on the pilot study that pertain to the fieldwork experience, methodological experiments and functionality of newly developed instruments, investigate whether proposed changes of the GGS questionnaire demonstrate methodological improvements and provide recommendations based on the results. The pilot study, which was conducted by the GGP team at the University of Ljubljana, is presented and documented in detail in deliverable D18 (Pilot Study Fieldwork Documentation), while the datasets of the pilot study are part of the deliverable D19 (Datasets of the pilot studies). The present deliverable builds on the previously mentioned deliverables and offers mainly an analytical report on the main issues, which were supposed to be empirically addressed by the pilot study:

1) What sort of experiences with the newly developed questionnaire do respondents report and how do these reports vary across different modes of data collection (face-to-face, telephone and web)?
2) What do analyses of methodological experiments tell us about the alternative solutions in the questionnaire? Do recommended solutions perform better than original ones in terms of measurement and questionnaire quality?
3) How do newly developed and revised scales perform in the pilot questionnaire in terms of psychometric properties? Are new solutions methodologically acceptable?

The results of these analyses will serve as an input for preparing the final version of the new GGS questionnaire (D26) and provide information for GGP Blueprint 2015 (D38). The report is structured in the following way: First, a short summary of the pilot study design and realization is presented, followed by the presentation of analyses and interpretations that refer to the duration of the questionnaire and some items which tapped respondent's experience with the survey. This is followed by a chapter on the analysis of five methodological experiments that were part of the pilot study. The last part of the report is an extensive set of analyses of the scales in the GGS survey that were renovated or newly developed and tested in the pilot study.

Before presenting the results of the analyses, some methodological clarifications are needed regarding the methods that we used. In general the evaluation of the items and item sets is based on a number of methods:

- Inspection of frequency distributions, means and standard deviations
- Inspection of correlations
- Reliability analyses
- Regression analyses
- Analyses of measurement equivalence

Reliability analysis was conducted through inspection of item correlations and Cronbach's alpha, which are standard procedures to gain insight into the internal consistency of scales (Hair et al., 1998; Hox, 2010). The type of regression analysis we performed was in most cases standard ordinary least squares linear regression, except in some cases where Poisson regression needed to be used due to the nature of distribution of the dependent variables (Cameron and Trivedi, 1998). Measurement equivalence was used to investigate the invariance of scales across different modes, where a structural equation modeling approach was used that allows to perform confirmatory factor analysis and to put equality constraints on various parameters under interest (Van den Schoot et al., 2012). It should be noted that we followed standard interpretation of fit indices in confirmatory factor analysis (Browne and Cudeck, 1993; Hu and Bentler, 1998).

One important methodological clarification refers to the suitability of Cronbach's alpha and confirmatory factor analysis (CFA) to evaluate the quality of measurement instruments. The CFA framework does not apply to all kinds of measurement instruments, but only to those which are essentially a scale (and not an index). Methods for estimating reliability and scale equivalence are based on an assumption that a battery of items under investigation forms a scale. A 'scale' refers to a measurement of a latent phenomenon, which is not directly observable and is only manifested through indicators. Each indicator measures the whole range of the phenomenon (min to max), but we use more indicators to ascertain structural validity and reliability (Hair et al., 1998). Some item-sets are clearly scales, since they are measuring latent phenomena and we can expect high correlations

between items (i.e. depression). Other sets are clearly an index (i.e. size of social network)and in these instances we do not compute reliabilities and perform CFA. For some item-sets we were not completely sure whether they are scales or indexes. In these instances, we decided to apply scaling methods anyway in order to investigate whether they indeed can be treated as scales.

2. Summary of the pilot study

In line with the decisions regarding the preparation of the pilot study design, the data collection occurred in two phases. The analyses in this report pertain to the first phase of data collection, which was intended to provide answers to the tasks that are central to the present deliverable. The second phase of data collection, the intention of which was to estimate the efficiency of mixed-mode design, is described in detail in the context of Deliverable D18.

The first phase of the pilot study implemented methodological experiments and a mode-effects design (for detailed information see D18). The survey was conducted on a sample of "panelists" who participated in a panel of the survey organization, and who were initially recruited in different ways (via telephone, face-to-face and using banners on different web portals and websites) to assure good sample structure. An important condition for the realization of the mode-effect design was that the respondents were reachable through all three contact modes (letter, telephone, e-mail). The data for the pilot was collected between the end of September and the middle of October 2011 in Slovenia on individuals aged 18 year and over. The starting sample (n=847) was randomly appointed to one of three modes (CATI, WEB, F2F). 621 respondents completed the survey, which makes a very good completion rate (73%), which varied to some extent across different modes (See table 1). The design included the use of incentives so that respondents received coupons in value of 5€.

Table 1 Survey final status – total sample

	F2F	CATI	WEB	Total
persons in initial sample	280	319	248	847
surveys completed	206	200	215	621
break off - never finished	0	8	24	32
break off - refused to finish	0	5	0	5
refused	18	5	0	23
completion rate	74%	63%	87%	73%

3. Experiences with the pilot

The respondents' experiences with the pilot were grasped with two types of data: one pertains to the duration of the questionnaire and the other to the attitudes toward the questionnaire. Length of the questionnaire and the related issue of duration of interviewing is an important aspect of respondent's burden and can be important factor of dropout and consequent unit nonresponse (Crawford et al., 2001). One of the major concerns of the GGP survey was its length. The following table (Table 2) presents average values of the duration across three different modes.

Table 2 Duration of the questionnaire across three different modes

Mode	Mean	N	Std. Deviation
FTF	0:52.24	206	0:13:36
CATI	1:02:17	200	0:14:21
WEB	0:55:31	215	0:23:31
Total	0:56:40	621	0:18:18

The duration was on average a little less than one hour, which is in the upper limit of the suggested length for surveys, especially for web and telephone administered surveys (Crawford et al., 2001; Sharp and Frankel, 1983). The shortest duration was for the F2F and the longest for CATI. The regression analysis of modes on the duration shows that WEB and CATI significantly influence the duration of the survey, with the impact of CATI being especially strong. This is reflected by the fact that CATI interviewing in average took more than an hour. This is probably also reflected in the dropout rate (see Table 1), which is largest for CATI. On the other hand, the duration and dropout do not seem to be related for

the WEB, which is somewhat surprising as the research shows that too long web surveys can seriously increase unit nonresponse (Crawford et al., 2001).

The second set of analyses pertains to the attitudes toward the questionnaires. These are items that appeared at the end of the questionnaire (a1202a to a1202e). Descriptive information on the answers provided is presented in Tables 3 and 5, whereas results of regression analyses that control for gender, age and level of education are presented in Tables 4 and 6.

Table 3 "Overall, how did you feel about completing this questionnaire?"

	Frequency	Percentage
1 very unpleasant	8	1%
2	12	2%
3	106	17%
4	304	49%
5 very enjoyable	191	31%
Total	621	100%

Table 4 Regression of modes on feelings toward the questionnaire

	b	s.e.	t	p
WEB	-.181	.080	-2.106	.023
CATI	-.257	.081	-2.347	.002
Sex	.047	.066	1.028	.480
Age	-.002	.003	-.938	.350
Education	-.039	.018	-1.986	.033

Weak mode effects can be noted, with the experience with the questionnaire being evaluated a bit less pleasant by respondents who answered the questionnaire in WEB and CATI mode in comparison to F2F mode. Despite this, respondents in general report positive experiences regardless the mode of data collection, as only 3% of them reported unpleasant or very unpleasant experiences.

Table 5 Attitudes toward the questionnaire

	Definitely not	2	3	4	Definitely yes
a1202a Was it difficult to answer the questions?	64.6%	17.4%	11.3%	5.7%	1.0%
a1202b Were the questions clear?	2.4%	6.9%	7.9%	26.2%	56.5%
a1202c Did the questions made you think?	12.6%	13.5%	24.3%	27.9%	21.7%
a1202d Was the topic interesting?	1.1%	1.9%	13.5%	34.6%	48.8%
a1202e Was the questionnaire too long?	40.1%	21.3%	18.9%	13.4%	6.3%

Table 6 Standardized regression coefficients of modes on attitudes toward the questionnaire (controlled for sex, age, education)

	A1202a	A1202b	A1202c	A1202d	A1202e
Web	0.13*	-0.01	0.22*	-0.04	0.03
CATI	0.09	-0.05	0.04	-0.14*	0.34*

note: * $p<0.01$

Attitudes toward the questionnaire were in general very positive. Although there were certain expectations that some parts of the questionnaire – especially complex event history batteries – would pose difficulties to respondents and that the length of questionnaire might be perceived as problematic, the results contradict those concerns to a large extent. Only a small proportion of respondents (6.7%) expressed that questions were difficult to answers, while 19.7% of respondents agreed that it was too long. Some mode effects can be noted, but these are not strong. The web respondents had a bit more negative attitude regarding the difficulty of the questionnaire, which is not surprising due to the self-administered nature of the questionnaire, which does not allow clarification of the questions. Of some concern is the result that CATI respondent were quite significantly more likely to complain about the length of the questionnaire.

To summarize, the presented results offer strong support for a claim that the respondent's experiences with the pilot survey were very good, even better than expected. Although the questionnaire is still rather long in comparison to typical surveys, only a small percentage of respondents found it unpleasant. We could conclude that the optimization of the

questionnaire was successful in terms of making individual items clear and easy to read and also in terms of the structure of flow, which increases respondents' interest for the topic. Although only for a minority of the respondents felt that the questionnaire is too long, it might be advisable to find options to further reduce its length.

4. Methodological experiments

The purpose of this section is to provide results of the analysis of several measurement experiments that were performed in the pilot study. These were conducted as *split-ballot* experiments in which the sample is split (roughly) in two experimental groups (Schuman and Presser, 1981). One group is administered one version of a particular questionnaire item (e.g. the original item wording in GGS wave 1) and the other group gets the other version (e.g. the renewed wording for the same item). Because the questionnaire was implemented in an electronic form, the practical implementation of split-ballot experiments was rather straightforward. A particular respondent was allocated to one experimental group or the other on the basis of his/her ID number. If the ID number was odd, for example, the respondent would be allocated to experimental group X; otherwise they would be allocated to experimental group Y. The following five split-ballot experiments were conducted in the context of the Slovenian GGS pilot:

- Scale alternatives for questions with a satisfaction scale (items a113, a217, a236, a263, a308, a311, a313, a418, a427, a701, a803, a824): One subsample had to answer satisfaction questions with scale values ranging from "Not at all satisfied" to "Completely satisfied" and the second subsample had to answer the same questions with scale values ranging from "Extremely dissatisfied" to "Extremely satisfied".
- Scale alternatives for items based on the Theory of Planned Behaviour (a278, a283, a413, a416, a425, a447, a614, a615, a616, a804, a815, a834): One subsample had to answer intention questions with a scale consisting of values 1-definitely not, 2-probably not 3-UNSURE, 4-probably yes, 5-definitely yes. A second subsample had to answer the same questions without the middle category 'unsure' and with an additional question 'Did you not provide an answer because you were unsure?' if no answer was provided to the main question.

- Ordering for questions on partner disagreements (a218, a219): One subsample was first presented with item a218, followed by item a219, while the second subsample first received item a219 and then item a218.
- Upper limit on the number of alters to be mentioned in response on a question about emotional social support (a501): One subsample had the possibility to enter an unlimited number of alters, while the other subsample had a limitation of max. 5 alters. This limitation was set by the system and was not mentioned in the question wording.
- Learning effect of name interpreters (a502): One subsample received a question a502, that asked for information on each of the network members mentioned at the first question of the network module, while the other subsample did not receive this question, but had to answer questions about network members at the end of the network module only.

It was undesirable that experimental groups would completely overlap, i.e. there would be a group of respondents who would have received original versions of all tested items while the other group would have received all renewed versions. If such a design were employed, the effect of conducting one experiment would have been confounded with the effects of others. Instead, the aim was to come as closely as possible to a balanced full factorial design in which every setting of one experimental factor appears with every setting of every other factor. There were thus actually 32 (25) versions of the questionnaire to which respondents were assigned at random. This setup assures that when analyzing the results of a single split-ballot experiment, we need only pay attention to how the respondents were split into two groups for that particular experiment and need not worry about possible contamination from other experiments that were conducted in the same questionnaire.

4.1 Satisfaction items

The questionnaire included a number of satisfaction items that pertained to various persons, activities and objects. A number of satisfaction items varied across respondents, as it was dependent on having a partner, child etc. We report the analysis of the following items:

a113: satisfaction with dwelling;
a217: satisfaction with relationship with partner;
a236_1: satisfaction with relationship with oldest child
a308_1: satisfaction with relationship with first mentioned HH member
a311: satisfaction with division of household tasks;
a313: satisfaction with division of childcare tasks;
a418 satisfaction with relationship with mother;
a427: satisfaction with relationship with father;
a701: satisfaction with life as a whole;
a803: satisfaction with current job.

All satisfaction items were part of a split ballot experiment in which the labeling of the endpoints of the 11-point answering scale varied. Version X has the labeling that was applied in previous waves of the GGS, whereas version Y has the labeling that is used in satisfaction items that are included in the European Social Survey. Below is an overview of all the items that were tested.

Table 7 Basic descriptive statistics of satisfaction items

	0	1	2	3	4	5	6	7	8	9	10	INR	DK	n	μ	σ
A113X	0.3	0.3	0.0	1.6	1.6	4.9	5.2	13.1	21.9	18.9	32.0	0.3	0.0	366	8.3	1.8
A113Y	0.0	1.2	0.4	2.0	1.2	5.1	6.3	13.3	28.6	18.8	22.4	0.4	0.4	255	8.0	1.8
A217X	0.4	0.0	0.7	0.7	0.7	1.4	2.5	5.3	23.0	19.9	45.4	0.0	0.0	282	8.8	1.5
A217Y	0.6	0.0	1.1	2.9	0.6	1.7	0.6	8.0	23.0	25.3	36.2	0.0	0.0	174	8.6	1.8
A236_1X	0.6	0.0	0.0	0.6	0.6	1.7	2.3	5.1	14.3	22.9	52.0	0.0	0.0	175	9.0	1.5
A236_1Y	1.0	0.0	0.0	0.0	0.0	2.9	1.0	8.7	12.5	28.8	44.2	0.0	1.0	104	8.9	1.5
A236_2X	1.6	0.0	0.0	0.0	0.0	0.0	2.4	8.7	12.7	27.0	46.8	0.0	0.8	126	9.0	1.6
A236_2Y	0.0	0.0	0.0	0.0	0.0	2.7	2.7	5.5	13.7	35.6	37.0	1.4	1.4	73	8.9	1.2
A263_1X	2.1	2.1	0.0	2.1	2.1	6.4	2.1	6.4	10.6	12.8	51.1	0.0	2.1	47	8.3	2.5
A263_1Y	2.6	0.0	0.0	2.6	2.6	2.6	2.6	5.3	5.3	31.6	39.5	0.0	5.3	38	8.5	2.3
A308_1X	0.9	1.7	0.0	0.9	0.9	4.3	1.7	9.6	27.8	16.5	34.8	0.0	0.9	115	8.3	2.0
A308_1Y	0.0	0.0	1.1	3.3	1.1	5.4	5.4	8.7	25.0	20.7	29.3	0.0	0.0	92	8.2	1.9
A311X	0.0	0.0	0.8	2.1	0.8	7.5	7.1	11.7	15.1	23.0	31.8	0.0	0.0	239	8.2	1.9
A311Y	0.7	0.0	1.4	2.9	3.6	3.6	5.0	11.5	20.9	23.7	26.6	0.0	0.0	139	8.0	2.1
A313X	0.0	0.0	0.0	2.3	3.4	3.4	3.4	12.5	14.8	21.6	38.6	0.0	0.0	88	8.4	1.8
A313Y	1.9	0.0	1.9	1.9	0.0	1.9	3.7	9.3	18.5	27.8	29.6	3.7	0.0	54	8.3	2.1

A418X	1.8	0.0	0.7	1.1	1.8	4.3	4.7	10.8	23.4	21.9	28.8	0.4	0.4	278	8.2	2.0
A418Y	0.5	1.4	0.5	3.2	3.7	2.8	4.1	11.9	16.1	23.4	32.1	0.5	0.0	218	8.1	2.1
A427X	2.8	1.4	1.9	3.7	4.2	5.6	5.6	9.7	22.7	16.7	25.0	0.9	0.0	216	7.5	2.6
A427Y	2.5	0.6	1.8	2.5	3.7	4.3	4.9	9.2	21.5	19.6	28.2	0.6	0.6	163	7.8	2.4
A701X	0.3	0.0	0.8	2.2	1.4	6.6	7.1	19.9	33.3	16.7	11.7	0.0	0.0	366	7.6	1.7
A701Y	0.8	0.0	0.4	1.6	1.6	9.0	5.5	18.8	28.6	22.0	11.4	0.4	0.0	255	7.6	1.8
A803X	1.6	0.0	1.6	1.6	0.0	4.8	4.8	8.1	9.7	8.1	59.7	0.0	0.0	62	8.6	2.2
A803Y	2.6	5.3	0.0	0.0	2.6	5.3	0.0	7.9	13.2	15.8	47.4	0.0	0.0	38	8.2	2.7
A824X	0.4	0.9	1.7	2.1	3.8	6.4	7.3	18.4	23.5	16.7	17.1	1.3	0.4	234	7.5	2.1
A824Y	2.5	0.6	2.5	1.9	3.1	4.4	5.6	12.5	31.2	22.5	10.6	2.5	0.0	160	7.4	2.2

The values in Table 7 suggest that there are no major differences between the two versions of the scales. Some variability is noted at the right extreme, where the percentages for the Y version are somewhat smaller than for the X version. To get insight into the question whether the type of scale significantly influences the satisfaction scores, linear regressions for each satisfaction variable were run.

In Table 8, for each of the satisfaction items, the regression effect for the scale version variable (dummy, where X version of the scale is the base category) is presented. The impact of the scale version is controlled for gender, age, education and mode.

Table 8 Regression effects of scale version on satisfaction score

Variable	b	s.e	t	p
Satisfaction with relationship with partner.	-.24	.16	-1.46	.15
Satisfaction with relationship with oldest child	-.10	.19	-.52	.61
Relationship with first mentioned household member	-.19	.27	-.71	.48
Satisfaction with division of household tasks	-.16	.21	-.79	.43
Satisfaction with division of childcare tasks	-.14	.35	-.40	.69
Satisfaction with relationship with mother	.02	.19	.13	.90
Satisfaction with relationship with father	.30	.27	1.11	.27
Satisfaction with life as a whole	.01	.14	.09	.93
Satisfaction with being retired or a homemaker	-3.38	1.22	-2.76	.01
Satisfaction with current job	-.06	.23	-.28	.78

The analysis makes clear that there are no statistically significant differences in the scores on these satisfaction variables, except for satisfaction with being retired or a homemaker, where the mean score for the Y version is significantly lower than for the X version. We additionally investigated whether mode has any effect on the satisfaction score, controlled for scale version, age, education and sex. The results are shown in Table 9.

Table 9 Regression effects of mode on satisfaction score

Variable	b	s.e	t	p
Satisfaction with dwelling				
CATI	-.15	.18	-.82	.41
WEB	-.60	.18	-3.35	.00
Satisfaction with partner relationship				
CATI	.05	.19	.24	.81
WEB	-.30	.19	-1.57	.12
Satisfaction with relationship with oldest child				
CATI	-.18	.23	-.79	.43
WEB	-.09	.22	-.39	.70
Satisfaction with relationship with first mentioned HH member				
CATI	-.54	.34	-1.59	.11
WEB	-.75	.34	-2.22	.03
Satisfaction with division of household labor				
CATI	.13	.25	.53	.60
WEB	-.20	.25	-.80	.42
Satisfaction with division of child care				
CATI	.57	.43	1.32	.19
WEB	.22	.42	.53	.60
Satisfaction with relationship with mother				
CATI	-.53	.23	-2.30	.02
WEB	-.42	.24	-1.80	.07
Satisfaction with relationship with father				
CATI	-.04	.32	-0.14	.89
WEB	-.48	.32	-1.49	.14
Satisfaction with life as a whole				
CATI	-.03	.17	-.19	.19
WEB	-.36	.17	-2.16	.03
Satisfaction with being retired or a homemaker				
CATI	-.29	.71	-.41	.68
WEB	-.98	.75	-1.31	.19
Satisfaction with current job				
CATI	.05	.28	.17	.87
WEB	-.60	.27	-2.22	.03

Investigation of Table 9 shows that for five satisfaction items we can note a statistically significant effect of WEB mode on the score, in all cases negative. The effect of CATI mode is present only for satisfaction with the relationship with a respondent's mother. It seems that the web respondents are inclined to give lower satisfaction scores in comparison to F2F, which might be the consequence of self-administration mode, which lowers the presence of social desirability effects. Answers in the self-administrated mode might be thus closer to the true value. Taking this into account, it seems that the majority of the satisfaction items are reasonably robust against mode effects

Additional tests (results not shown) make clear that the standard deviation of the balanced version (Y) is larger for the question of satisfaction with the partner relationship ($p < .02$), and also larger – but not statistically significant – for division of household labor ($p < .11$) and division of child care ($p < .13$). In addition, it was also tested whether the version effect differed across modes for satisfaction with the partner relationship, but it did not. In combination with the fact that Y is balanced, this suggests that the Y version is methodologically doing slightly better.

We recommend using satisfaction questions with the end-labels running from 'extremely dissatisfied' to 'extremely satisfied', preferably in a self-administered mode.

4.2 Theory of planned behavior items

The intention of this experiment was to test the recommendation to replace the original 4-point response scale for measuring intentions (definitely not, probably not, probably yes, definitely yes) with a 5-point scale by adding a mid-point 'unsure'. The rationale behind this recommendations are: 1) 5-point scale gives R an opportunity to indicate they are genuinely unsure, i.e., unsure is a valid response to questions about intention; 2) 5-point scale has greater variance; 3) Midpoint of 5-point scale does not differ significantly from midpoint of 4-point scale; 4) 5-point scale is associated with higher predictability of intentions.

All the tested items inquired into how likely it would be that the respondent does one of the following (within the next three years):

a278: start living with partner;
a283: marry partner;
a413: start living with parents;
a416: start living with mother;
a425: start living with father;
a447: start living separately from parents;
a614: have a/another child;
a615: adopt a child;
a616: have any (more) children at all;
a804: take a job;
a815a: resume work after leave has ended;
a815b: intend to work after leave has ended;
a834: retire

The 5-point scale is denoted as X, while the 4-point scale is denoted as Y. In the latter case, if the respondent did not provide the answer to a question, an additional question was displayed, inquiring into the cause for the item nonresponse, e.g.

"You did not provide an answer to the previous question: '*Do you intend to start living with your partner during the next 3 years?* 'Did you skip this question because you were unsure of the answer or for some other reason?"

Two options ("I was unsure" and "other reason") were offered for such respondents. The number of item nonresponses in the pilot was very low, however. The aforementioned additional question was administered only to two respondents: one after question a413Y and the other after a614Y. In both cases the answer was "I was unsure".

Table 10 Row percentages for intention items

	Definitely not	Probably not	Unsure	Probably yes	Definitely yes	Total n
a278X	13.7	24.4	20.6	27.5	13.7	131
a278Y	9.8	27.7		40.2	22.3	112
a283X	33.0	27.7	22.0	11.5	5.8	191

a283Y	32.7	39.4		23.0	4.8	165
a413X	69.5	22.9	3.8	1.0	2.9	105
a413Y	69.2	26.9		2.9	1.0	104
a416X	59.5	30.4	5.1	0.0	5.1	79
a416Y	71.0	27.4		0.0	1.6	62
a425X	81.2	12.5	3.1	0.0	3.1	32
a425Y	71.9	25.0		3.1	0.0	32
a447X	13.1	20.2	16.7	28.6	21.4	84
a447Y	13.7	27.4		30.1	28.8	73
a614X	43.7	22.8	12.2	11.0	10.3	263
a614Y	42.1	24.7		22.6	10.6	235
a615X	75.8	19.1	3.9	1.2	0.0	335
a615Y	67.8	29.7		2.1	0.3	286
a616X	63.6	17.8	3.3	8.3	7.0	242
a616Y	61.0	22.9		5.4	10.7	205
a804X	73.0	7.9	9.5	4.8	4.8	63
a804Y	75.7	18.9		5.4	0.0	37
a815aX	0.0	0.0	0.0	20.0	80.0	5
a815aY	0.0	0.0		0.0	100.0	4
a815bX	0.0	0.0	0.0	0.0	100.0	1
a815bY	0.0	0.0		0.0	100.0	1
a834X	76.4	23.6	0.0	0.0	0.0	55
a834Y	71.1	23.7		2.6	2.6	38

Table 10 shows that the highest percentages of »unsure« category were selected for a278 (intend to start living with partner), **a283** (intend to marry partner), **a447** (intend to start living separately from parents) and **a614** (intend to have a/another child).

The data from the TPB split ballot experiment were further analyzed in two ways. First, by regarding the data as nominal, a chi-squared test is performed. And secondly, by recoding and regarding the data as numeric, a linear regression can be applied.

The chi-squared test was performed by disregarding the “unsure” category of the first scale. All units falling within this category were discarded. The test was performed on the remaining 2 by 4 table (scale type by answer category). For each TPB item, the table of observed frequencies was compared to the table of expected frequencies. The expected

frequencies were calculated from the margins, assuming the two cross-tabulated quantities were independent (this is the most common chi-squared setup).

The null hypothesis, therefore, states that the distribution of answers into the "definitely not", "probably not", "probably yes", and "definitely yes" categories is identical for the two versions of the scale. In other words, under the null hypothesis, introducing the middle ("unsure") category may draw some respondents to choose this category, but the distribution of answers is unchanged for the remaining categories (the ratio of "definitely not" to "probably not" remains unchanged etc.).

In order to perform the chi-square test, the category "probably yes" was omitted for item a416, because it was empty for both scale types. The number of respondents providing an answer to item a815a and a815b was too low to perform the test.

Table 11 Chi-squared test for TPB items

Variable	chi-sq	df	p
a278	3.55	3	0.31
a283	5.85	3	0.12
a413	2.27	3	0.52
a416	1.88	2	0.39
a425	3.50	3	0.32
a447	0.55	3	0.91
a614	8.30	3	0.04
a615	10.30	3	0.02
a616	4.55	3	0.21
a804	3.83	3	0.28
a834	2.98	3	0.39

The figures in Table 11 show that the null hypothesis can be rejected at .05 level for items a614 (intention to have a/another child) and a615 (intention to adopt a child). The results therefore indicate that adding a middle category changes the distribution of respondents' answers into the remaining four categories for these two items only. Let us note, however, that when a large number of tests (in our case 11 chi-squared tests) are performed with no

prior hypotheses, it is possible that a few tests turn out significant purely by capitalizing on chance (the problem known as *multiple testing*).

The regression analysis presented in this section makes the assumption that intention items can be regarded as interval variables, although strictly speaking they are more of an ordinal nature. This assumption is made in order to perform linear regression that has a higher statistical power than the chi-squared test. Another benefit of linear regression is the ease of controlling for potential confounding variables (like mode of administration). Each intention item was regressed onto scale version, mode of administration and basic demographic characteristics of the respondent. In Table 12, we report effects of scale version for all intention items.

Table 12 Regression effects of scale version on intention items

	b	s.e	t	p
a278	0.28	0.16	1.73	0.09
a283	-0.06	0.13	-0.46	0.64
a413	-0.05	0.11	-0.51	0.61
a416	-0.28	0.15	-1.89	0.06
a425	0.08	0.19	0.42	0.68
a447	0.08	0.23	0.36	0.72
a614	0.10	0.12	0.82	0.41
a615	0.03	0.05	0.69	0.49
a616	-0.07	0.09	-0.81	0.42
a804	-0.61	0.32	-1.92	0.06
a834	0.24	0.14	1.76	0.08

The scale effect is not statistically significant at .05 level for any of the intention items. These results are congruent with the analysis of chi-square tests and suggest that no effect of scale version is demonstrated for intention items. The two different types of analysis might suggest that there might be a weak effect for certain items (significant at 0.10 level), but the items identified do not coincide with chi-square tests. The multiple testing problem makes it even harder to claim that the weak patterns that were revealed are of any substance. We cannot favor either version of the scale on the basis of the analyses in this section.

Due to the theoretical rationale for the 5-point scale (see above) it is recommended that 4-point scale, used in GGS wave 1 should be replaced with a 5-point scale that includes a midpoint 'unsure'.

4.3 Ordering of household disagreement items

The next split-ballot experiment concerned the ordering of items a218 and a219 in the questionnaire. In the first version, respondents were first asked about the frequency of disagreements about different topics with their partner (a218) and then about the ways they deal with serious disagreements (a219). In the second version, the order was reversed. The original ordering is denoted with X, while the reverse ordering is denoted with Y in the following analyses.
Battery a218 included the following topics that couples may disagree on:

a218a: household chores,
a218b: money,
a218c: use of leisure time,
a218d: relations with friends,
a218e: relations with parents and in-laws,
a218f: having children,
a218g: child-raising issues.

Battery a219 included the following ways of dealing with serious disagreements:

a219a: avoid discussion by giving in;
a219b: discuss your disagreement calmly;
a219c: argue heatedly or shout;
a219d: refuse to talk about it.

The scale for both questions ranged from "never" to "very frequently". Both the chi-squared test and linear regression were again used to analyze the data.

Table 13 Row percentages for items a218

	never	seldom	sometime	freq.	very freq.	INR	not appl.	Total n
a218aX	25.4	31.0	31.5	9.9	1.3	0.9		232
a218aY	34.4	41.5	18.3	3.6	1.3	0.9		224
a218bX	28.0	33.2	26.3	9.9	0.4	2.2		232
a218bY	29.9	40.2	19.6	7.6	1.8	0.9		224
a218cX	28.4	37.5	22.4	7.8	1.7	2.2		232
a218cY	25.4	37.1	23.7	8.9	3.6	1.3		224
a218dX	47.8	34.5	10.3	5.2	0.0	2.2		232
a218dY	43.3	39.3	9.4	4.9	1.8	1.3		224
a218eX	45.7	29.7	13.8	6.5	2.2	2.2		232
a218eY	44.6	26.8	18.8	6.7	0.9	2.2		224
a218fX	84.9	7.8	3.4	1.3	0.0	2.6		232
a218fY	80.8	10.3	5.4	2.2	0.0	1.3		224
a218gX	37.9	25.0	16.8	5.2	1.3	2.2	11.6	232
a218gY	34.4	25.0	18.8	8.0	0.4	1.3	12.1	224

Table 14 Row percentages for items a219

	never	seldom	sometime	freq.	very freq.	INR	not appl.
a219aX	19.9	31.7	33.0	10.0	4.1	1.4	221
a219aY	17.9	21.0	42.4	15.2	3.1	0.4	224
a219bX	0.0	7.2	23.1	49.3	19.0	1.4	221
a219bY	0.9	5.8	17.9	49.6	24.1	1.8	224
a219cX	29.4	34.4	25.3	8.6	0.5	1.8	221
a219cY	31.7	37.9	20.5	6.2	1.3	2.2	224
a219dX	37.1	40.3	15.8	4.1	0.9	1.8	221
a219dY	37.1	33.5	21.4	4.9	0.4	2.7	224

The chi-squared test was applied to 5 by 2 (answer category by experimental treatment) tables for each item separately. The category “very frequently” was omitted for item a218f, because if was empty for both experimental treatments.

Table 15 Chi-squared test for items a218 and a219

	chi-sq	df	sig.
a218a	21.16	4	0.00
a218b	6.44	4	0.17
a218c	2.12	4	0.71
a218d	5.49	4	0.24
a218e	3.30	4	0.51
a218f	2.53	3	0.47
a218g	2.99	4	0.56
a219a	10.36	4	0.03
a219b	5.15	4	0.27
a219c	3.50	4	0.48
a219d	3.77	4	0.44

Table 15 exhibits a rather clear pattern: the reversal of the ordering of questions a218 and a219 has the highest impact on the first item in each battery. The second item in the battery is less affected and so forth.

Table 16 Regression effects of scale version on scores of a218 and a219

Variable	b	s.e	t	p
a218 household chores	-0.34	0.09	-3.87	0.00
a218b money	-0.11	0.09	-1.25	0.21
a218c use of leisure time	0.14	0.10	1.46	0.15
a218d relations with friends	0.08	0.08	0.99	0.32
a218e relations with parents and in-laws	0.01	0.09	0.15	0.88
a218f having children	0.08	0.06	1.44	0.15
a218g child-raising issues	0.09	0.10	0.84	0.40
a219a avoid discussions by giving in	0.18	0.10	1.82	0.07
a219b discuss your disagreement calmly.	0.13	0.08	1.65	0.10
a219c argue heatedly or shout.	-0.10	0.09	-1.08	0.28
a219d refuse to talk about it.	0.05	0.09	0.60	0.55

The results of the regression analyses are similar to those of the chi-squared tests (see Table 16). The first topic (disagreements about household chores) is reported to be more common when topics of disagreements are asked first. Analogously, the first two ways of dealing (giving in and discussing calmly) are reported as more common when ways of dealing with serious disagreements are asked first. The respondents therefore seemingly report a higher frequency when they first encounter either battery.

Another possible interpretation that also takes into account the substance of the questions is the following. When ways of dealing with serious disagreements are asked first, the respondent finishes the first battery of questions by evaluating the frequency of rather unconstructive ways of dealing with disagreements of "argue heatedly or shout" and "refuse to talk about it". With this in mind, the respondent continues to evaluate the frequency of potential topics of disagreement and evaluates that such unconstructive ways of dealing with disagreements are not often employed. This effect fades with subsequent topics of disagreement.
The recommendation is to keep the current ordering of the questions in order to i) assure comparability with previous waves and ii) avoid priming the respondents as to what kind of disagreements we have in mind.

Next, mode effects were examined for both sets of items. Results of regression analyses are presented in Table 17. Mode effects are observed in case of WEB for four (out of seven) disagreement items and none for solving tactics, while in case of CATI mode effect is observed for three disagreements items and one for solving tactics. Respondents mention more disagreements in the WEB version than in the CATI and CAPI version. ***It is recommended to present these items in a self-administered mode in CAPI in order to minimize mode effects between CAPI and WEB.***

Table 17 Regression effects of mode on scores of a218 a 219

Variable	b	s.e	t	p
a218a household chores				
CATI	-0.31	0.11	-2.85	0.00
WEB	0.22	0.11	2.05	0.04
a218b money				

CATI	-0.12	0.11	-1.04	0.30
WEB	0.32	0.11	2.81	0.01
a218c use of leisure time				
CATI	-0.24	0.12	-2.08	0.04
WEB	0.18	0.12	1.54	0.12
a218d relations with friends				
CATI	-0.08	0.10	-0.75	0.45
WEB	0.21	0.10	2.07	0.04
a218e relations with parents and in-laws				
CATI	0.17	0.12	1.48	0.14
WEB	0.40	0.12	3.45	0.00
a218f having children				
CATI	-0.00	0.07	-0.04	0.97
WEB	0.11	0.07	1.61	0.11
a218g child-raising issues				
CATI	0.07	0.13	0.52	0.60
WEB	0.23	0.12	1.86	0.06
a219a avoid discussions by giving in				
CATI	-0.10	0.12	-0.84	0.40
WEB	0.12	0.12	1.04	0.30
a219b discuss your disagreement calmly.				
CATI	0.12	0.10	1.17	0.24
WEB	-0.09	0.10	-0.88	0.38
a219c argue heatedly or shout.				
CATI	-0.13	0.11	-1.17	0.24
WEB	0.12	0.11	1.03	0.31
a219d refuse to talk about it.				
CATI	-0.23	0.11	-2.13	0.03
WEB	-0.03	0.11	-0.26	0.79

Negligible mode effects were observed for the set of items on how respondents *solve disagreements* with their partner (2.19). *It is recommended to retain this set of items and pose them after the set on partnership disagreements.*

4.4 Cap or no cap in social network questions?

Questionnaire item a501 was a name generator that requested the respondent to list those persons with whom he/she discussed important things over the last 12 months. It was implemented in two versions:

- X: the respondent could name an unlimited number of alters
- Y: the maximum number of alters the respondent could name was five

The intention of this experiment was to investigate whether the limitation of number of possible alters has any effect on the size of the emotional support network. The limitation to max. 5 alters was set by the system and was not mentioned in the question wording. We applied two methods to investigate this question: one is regression analysis on the size of social network and the other is investigation of the structure of named alters.

The method of analysis that is applied to count data such as these is *Poisson regression*. As Table 18 shows, the shape distribution of the distribution is typical Poisson. The results of Poisson regression in Table 19 can be interpreted analogously to those of linear regression. The method merely applies the appropriate distributional assumption to correctly evaluate the standard errors of the effects.

Table 18 Row percentages for number of alters named at name generator a501

	0	1	2	3	4	5	6	7	8	9	10	11	n	μ
F2F no cap	2.0	41.0	22.0	17.0	10.0	3.0	2.0	0.0	2.0	1.0	0.0	0.0	100	2.28
F2F capped	6.6	32.1	41.5	8.5	6.6	4.7							106	1.91
CATI no cap	9.9	25.7	21.8	19.8	9.9	6.9	4.0	0.0	0.0	1.0	0.0	1.0	101	2.47
CATI capped	5.1	16.2	17.2	26.3	20.2	15.2							99	2.86
WEB no cap	9.7	38.8	17.5	12.6	10.7	3.9	2.9	2.9	1.0	0.0	0.0	0.0	103	2.19
WEB capped	4.5	36.6	17.0	11.6	13.4	17.0							112	2.44

Table 19 Poisson regression for item a501: discussing important things

	b	s.e.	z	p
Version	-0.21	0.10	-2.14	0.03
CATI	0.08	0.09	0.90	0.37
WEB	-0.11	0.09	-1.19	0.23
Sex	0.23	0.05	4.28	0.00
Education	0.05	0.19	0.28	0.78
age	-0.01	0.00	-6.28	0.00
Version X CATI	0.26	0.13	1.96	0.05
Version X WEB	0.34	0.13	2.54	0.01

The interaction term between the mode of administration and the name generator version is significant. For F2F administration, the limit to five alters results in less alters named, as would be expected. The results for the other two modes are somewhat surprising as the effect is significant, but in the opposite direction. The reason for this is unclear as the respondents were not informed about the limitation to five alters unless they reached it. To be better able to interpret the results the structure of alters was investigated.
The alters that were named at a501x and a501y are examined. The structure of the alters' sex, age and relation to respondent is presented in Tables 20 through 23. In each table, two types of percentages are given:

- A-type percentages refer to the proportion of alters that fall within a certain category. The problem with regarding the alter as the unit of analysis is that respondents who named more alters have a larger bearing on the results.
- B-type percentages aim to correct this and refer to the number of all respondents who named at least one alter that falls into the given category. Note that B-type percentages need not sum up to 100.

Table 20 Composition of alters' sex for alters named at item a501x (no cap)

	F2F		CATI		WEB		Total	
2-3 4-5 6-7 8-9	A	B	A	B	A	B	A	B
male	33.3	47.0	31.3	50.5	32.3	51.5	32.3	49.7
female	66.7	85.0	68.7	78.2	65.9	68.9	67.1	77.3
INR	0.0	0.0	0.0	0.0	1.8	3.9	0.6	1.3
total n (alters—egos)	228	100	249	101	226	103	703	304

Table 21 Composition of alters' sex for alters named at item a501y (capped)

	F2F		CATI		WEB		Total	
2-3 4-5 6-7 8-9	A	B	A	B	A	B	A	B
male	37.6	55.7	33.2	59.6	26.7	41.1	32.1	51.7
female	62.4	72.6	66.8	85.9	72.2	84.8	67.5	81.1
INR	0.0	0.0	0.0	0.0	1.1	2.7	0.4	0.9
total n (alters—egos)	202	106	283	99	273	112	758	317

Table 22 Composition of alters' age for alters named at a501x (no cap)

	F2F		CATI		WEB		Total	
2-3 4-5 6-7 8-9	A	B	A	B	A	B	A	B
-15	0.4	1.0	0.0	0.0	0.4	1.0	0.3	0.7
15-25	17.5	20.0	9.6	15.8	23.9	24.3	16.8	20.1
25-35	23.2	37.0	24.5	38.6	32.3	43.7	26.6	39.8
35-45	23.2	33.0	22.5	37.6	14.6	28.2	20.2	32.9
45-55	14.0	27.0	18.9	34.7	13.7	20.4	15.6	27.3
55-65	12.7	24.0	16.5	28.7	9.7	16.5	13.1	23.0
65+	6.6	12.0	6.8	12.9	2.2	4.9	5.3	9.9
INR	2.2	3.0	1.2	3.0	3.1	5.8	2.1	3.9
total n (alters—egos)	228	100	249	101	226	103	703	304

Table 23 Composition of alters' age for alters named at a501y (capped)

	F2F		CATI		WEB		Total	
2-3 4-5 6-7 8-9	A	B	A	B	A	B	A	B
-15	0.0	0.0	0.7	2.0	0.0	0.0	0.3	0.6
15-25	12.9	15.1	19.1	27.3	16.8	21.4	16.6	21.1
25-35	24.3	32.1	22.6	42.4	32.2	48.2	26.5	41.0
35-45	20.3	30.2	18.7	35.4	22.3	38.4	20.4	34.7
45-55	20.3	34.0	21.2	39.4	12.5	25.9	17.8	32.8
55-65	13.4	21.7	10.6	22.2	12.1	23.2	11.9	22.4
65+	7.4	13.2	4.9	13.1	2.6	6.2	4.7	10.7
INR	1.5	1.9	2.1	4.0	1.5	3.6	1.7	3.2
total n (alters—egos)	202	106	283	99	273	112	758	317

Capping item a501 at five alters —rather surprisingly— did not have the effect of reducing the number of named alters. The demographic structure of alters was also very similar for both treatment conditions.

The recommendation is to use the uncapped version.

4.5 Learning effect of item a502 in the context of social network questions

The intention of this split-ballot experiment was to investigate whether a follow-up question to the name generator for emotional social network produces a learning effect,

resulting in fewer named alters in the following name generator questions. More specifically, half of the respondents in the sample (experimental group Y) received a follow-up question to item a501 (name generator of emotional support network) inquiring into the means (face-to-face, video conference, phone etc.) of discussing important matters for *each* named alter. Control group X received no such follow-up question. It is expected that experimental group Y learned that it might be wiser to name fewer alters since they might be asked follow-up questions for each alter named. If so, this would be reflected in a lower number of alters named in the rest of name generators in the module as compared to the baseline group X that did not receive question a502. Note that the answer "Me" (code 99) on items a503 and a513 was disregarded, i.e. not counted as an alter. Descriptive analysis of the name generator items for both groups of respondents is first presented.

Table 24 Row percentages for number of alters named at network questions

	0	1	2	3	4	5	6+	n	μ
a503X	0.0	37.5	18.8	25.0	6.2	6.2	6.2	16	2.50
a503Y	4.3	21.7	34.8	26.1	13.0	0.0	0.0	23	2.22
a505X	2.3	40.9	36.4	6.8	9.1	4.5	0.0	44	1.93
a505Y	1.8	23.6	30.9	25.5	12.7	0.0	5.5	55	2.53
a511X	72.1	20.8	6.2	0.6	0.3	0.0	0.0	308	0.36
a511Y	64.5	22.0	9.9	2.9	0.3	0.3	0.0	313	0.53
a513X	1.3	54.3	32.7	8.1	2.7	0.4	0.4	223	1.60
a513Y	1.3	49.6	35.1	10.1	3.5	0.4	0.0	228	1.66
a515X	88.0	8.4	1.9	1.0	0.6	0.0	0.0	308	0.18
a515Y	87.5	8.9	3.2	0.0	0.0	0.3	0.0	313	0.17
a519X	84.7	11.0	3.6	0.6	0.0	0.0	0.0	308	0.20
a519Y	83.7	11.8	1.9	1.6	0.6	0.3	0.0	313	0.25
a522X								0	
a522Y	0.0	100.0	0.0	0.0	0.0	0.0	0.0	1	1.00
a524X								0	
a524Y	0.0	100.0	0.0	0.0	0.0	0.0	0.0	1	1.00
a528X	89.6	8.1	1.9	0.3	0.0	0.0	0.0	308	0.13
a528Y	91.7	6.1	1.0	1.3	0.0	0.0	0.0	313	0.12
a530X	89.3	6.8	3.2	0.0	0.3	0.0	0.3	308	0.17
a530Y	90.7	5.1	2.6	1.0	0.6	0.0	0.0	313	0.16
a532X	90.6	6.2	2.3	1.0	0.0	0.0	0.0	308	0.14
a532Y	92.0	5.8	1.9	0.3	0.0	0.0	0.0	313	0.11

a535X	80.5	16.6	2.9	308	0.22
a535Y	75.4	21.7	2.9	313	0.27

Table 24 does not offer much evidence of learning effect, therefore we performed a regression analysis of questionnaire version (a502 asked or not) on the social network items. Results are presented in Table 25.

Table 25 Regression effects of questionnaire version on number of listed alters (controlled for mode, sex, age, education)

Variable	b	s.e	t	p
a503: receiving support with childcare, HH members	-0.04	0.23	-0.19	0.85
a505: receiving support with childcare, other	-0.32	0.14	-2.28	0.02
a511: giving support with childcare	-0.37	0.12	-3.02	0.00
a513: receiving support with HH tasks, HH members.	-0.06	0.07	-0.82	0.41
a515: receiving support with HH tasks, other	0.01	0.19	0.03	0.97
a519: giving support with HH tasks	-0.25	0.17	-1.43	0.15
a528: giving support with personal care.	0.07	0.23	0.29	0.78
a530: receiving financial support.	-0.60	0.41	-1.46	0.15
a532: giving financial support.	0.29	0.23	1.26	0.21
a535: receiving inheritance.	-0.16	0.16	-0.96	0.34

The data were analyzed with Poisson regression. The results indicate that a learning effect might be present: the effect of asking item 502 on the number of alters is negative and significant for items a505 and a511. One possible interpretation is that experimental group Y who received the additional question was careful not to name too many alters at these name generators. As no more follow-up questions were being posed, the learning effect faded.

Item a503 was the first item to follow a502 but does not exhibit the learning effect. This, however, could be explained with the more factual nature of this question that inquiries into childcare support received from other household members. This may be less open to

interpretation than receiving and giving childcare support to other people, hence the negligible difference between experimental groups X and Y for item a503.
Because the additional questionnaire item clearly had an effect on the number of alters named at subsequent name generators, the recommendation is to exclude it.

5. Functionality of newly developed instruments

This part of the report focusses on the analysis of measurement scales that were renewed for the pilot study. The issue of measurement quality was approached from the perspective of scale distribution, reliability in terms of internal consistency, structural validity and scale equivalence. Standard descriptive statistics were computed to investigate scale distribution (frequency distributions, means and standard errors). Internal consistency was assessed on the basis of Cronbach's alpha, where the rule of thumb is that an alpha value of 0.6–0.7 indicates acceptable reliability and that a value of 0.8 or higher indicates good reliability (Brown, 2006). Structural validity is one dimension of construct validity and is estimated by means of confirmatory factor analysis, demonstrating the extent to which a proposed measurement model fits the data (Bollen, 1989; DeVellis, 2010). To investigate the mode effects on scales, an equivalence approach was undertaken, searching for metric and scalar invariance across different mode groups. While metric equivalence pertains to the invariance of the factor loadings between items and theoretical constructs, scalar equivalence pertains to equality constraint of the intercepts across different groups (Van de Vijver, 1998).

5.1 Household-related activities

5.1.1 Division of household labor

The number of items in this set was reduced compared to earlier waves. The goal was to obtain a set with some items being more often performed by the female partner and some being more often performed by the male partner. The current version includes the following items:

a310a: preparing daily meals
a310b: vacuum-cleaning the house
a310c: doing small repairs
a310d: paying bills and keeping financial records
a310e: organizing joint social activities

These items were recoded, so that a score of 1 means 'always by the male partner' and a score of 5 'always by the female partner'. The answer 'always or usually some else' was given by 5% or less of the respondents, and is disregarded in the following analyses.

Table 26 Descriptive statistics of the division of household labor items

Variables	N	Mean	SD	Min	Max
a310a: preparing daily meals	360	3.65	0.94	1	5
a310b:vacuum-cleaning the house	356	3.44	1.08	1	5
a310c: doing small repairs	361	2.04	1.01	1	5
a310d: paying bills and keeping financial records	368	3.01	1.27	1	5
a310e: organizing joint social activities	366	3.25	0.77	1	5

Table 26 shows that three of the five tasks are –on average– more often said to be performed by the female partner (preparing daily meals, vacuum-cleaning the house and organizing joint social activities). One task is said to be performed more often by the male partner (doing small repairs in and around the house). Finally, one task is –on average– said to be performed equally often by both partners (paying bills and keeping financial records). All of the five items have only weak correlations with the others (highest correlation is .18 between cooking and vacuum-cleaning), so they do not form a scale. They could, however, still be viewed as an index.

Table 27 Regression effects of mode on division of household labor (controlling for gender, age, educational attainment, partner status and version)

Variable	b	s.e	t	p
a310a: Preparing daily meals				
CATI	-.17	.12	-1.44	.15
WEB	-.22	.12	-1.82	.07

a310b: *Vacuum-cleaning the house*				
CATI	.24	.14	1.72	.09
WEB	-.14	.14	1.00	.32
a310c: *Doing small repairs in and around the house*				
CATI	.19	.13	1.46	.15
WEB	.37	.13	2.90	.01
a310d: *Paying bills and keeping financial records*				
CATI	-.01	.15	-.03	.97
WEB	.12	.15	.78	.43
a310e: *Organizing joint social activities*				
CATI	.02	.09	.21	.83
WEB	-.04	.10	-.40	.69

The mode differences are generally relatively modest and non-significant, with some exceptions (see Table 27). CATI respondents are more likely to report that the female partner is doing the vacuum-cleaning than WEB respondents. WEB respondents are more likely to report that the female partner is doing small repairs than CAPI respondents. One conclusion might be that this set of five items works fine from a substantive point of view.
Our recommendation regarding this set of items is to retain it as it is.

5.1.2 Division of child care labor

The number of items in this set was also reduced compared to earlier waves. Those items that showed the least variation were dropped. The current version includes the following items:

a310a: dressing the children or seeing that the children are properly dresses
a310b: staying at home with the children, when they are ill
a310c: playing with the children
a310d: helping the children with homework

These items were recoded, so that a score of 1 means 'always by the male partner' and a score of 5 'always by the female partner'. The answer 'always or usually some else' or

'children do it themselves' was given by a maximum of 10% or less of the respondents, and is disregarded in the following analyses.

Table 28 Descriptive statistics of the division of household labor items

variables	N	Mean	SD	Min	Max
a310a: dressing the children or seeing that the children are properly dresses	131	3.83	0.88	1	5
a310b:staying at home with the children, when they are ill	132	3.86	1.01	1	5
a310c: playing with the children	140	3.13	0.43	1	5
a310d: helping the children with homework	105	3.31	0.85	1	5

Table 28 shows that all of the tasks are – on average – more often said to be performed by the female partner. The gender differences are largest for 'dressing the children' and 'staying at home when the children are ill', and much smaller for the other two items. All correlations between the four items are positive and vary between .21 and .56. Cronbach's α is. 63, which is low, but may still be viewed as acceptable given that there are only four items. Reliability is lower in the CATI mode (α = .52) than in CAPI mode (α = .63) or WEB mode (α = .66).

Table 29 Regression effects of mode on division of child care labor (controlling for gender, age, educational attainment, partner status and version)

Variable	b	s.e	t	p
a310a: *Dressing the children*				
CATI	-.07	.20	-.36	.72
WEB	-.04	.19	-.23	.82
a310b: *Staying at home with the children when they are ill*				
CATI	.03	.23	.13	.89
WEB	-.13	.22	-.56	.58
a310c: *Playing with the children*				
CATI	-.11	.10	-1.17	.24
WEB	-.11	.09	-1.15	.25
a310d: *Helping the children with homework*				
CATI	.31	.22	1.39	.17
WEB	.10	.21	.47	.64

As can be observed from Table 29, the mode differences are modest and none of them come close to statistical significance. **Therefore, our recommendation is to retain this set of items as it is.**

5.1.3 Household decision-making

This set of questions on household decision-making was posed in exactly the same way in earlier waves of the GGS and it is only analyzed for possible mode effects. The scale is composed of the following items:

a312a: routine purchases for the household
a312b: occasional more expensive household purchases
a312c: the time you spend in paid work
a312d: the time your partner spends in paid work
a312e: the way children are raised
a312f: social life and leisure activities

Answers were recoded so that 1 means 'always the male partner ' and 5 means 'always the female partner'. Items d. and e. were recoded, so that one item is on the time that the wife spends in paid work and the second item on the time that the husband spends in paid work. In Table 30, the mode differences in the answer patterns for each of the six items are presented.

Table 30 Regression effects of mode on household decision-making (controlling for gender, age, educational attainment, partner status and version)

Variable	b	s.e	t	p
a312a: Routine purchases				
CATI	-.18	.12	-1.54	.13
WEB	-.05	.12	-.45	.17
a312b: Expensive purchases				
CATI	-.10	.09	-1.18	.24
WEB	-.15	.09	-1.68	.09
a312c: Time female partner spends in paid work				
CATI	.16	.13	1.21	.23

WEB	-.37	.13	-2.89	.01
a312d: Time male partner spends in paid work				
CATI	.06	.13	.46	.64
WEB	.51	.13	4.09	.01
a312e: Way children are raised				
CATI	.11	.07	1.45	.15
WEB	.02	.07	.22	.83
a312f: Social life and leisure activities				
CATI	-.03	.06	-.56	.57
WEB	-.00	.06	-.06	.95

Mode differences are small or non-existing for four of the six items. However, there are clear mode differences between WEB on the one hand and CATI and CAPI on the other hand concerning decision-making on paid work. In WEB mode, respondents are less likely to say that the decision is always made by the person that is concerned, i.e. that the male partner decides on his working hour by himself and the female partner decides on her working hours by herself. It is unclear how to interpret this finding. It could suggest that respondents in WEB mode experience more freedom to deviate from a norm that each partner separately decides on his or her working hours. However, one could also argue that working hours is an issue that affects both partners and this would make it an issue for joint decision-making.

Recommendations:

- **It is recommended to drop the item 'social life and leisure activities' (a312f). The variation is very low. In addition, it is unclear what joint decision means. It does not have to imply that the activities themselves are also jointly undertaken. Therefore, the substantive meaning of the question is relatively limited.**
- **It is recommended to retain the five other items.**

5.2 Networks

In this part, the analysis of all network-type items is presented. These are so called egocentric network questions, which ask respondents to list people (alters), who are

specifically related to the respondent (Kogovšek and Ferligoj, 2005). The network items are composed of a series of 'name generators' and one name interpreter, which asks respondent to define for each alter his/her sex, birth date or age, activity, relationship to the respondent and respondent's satisfaction with the relationship with alter. In the GGS questionnaire egocentric network type items appear in the household composition section (items on other household members) and in the network delineation and support module.

5.2.1. Other household members

In order to examine the possible effect of the mode of administration on the number of alters named on items a301b and a302b, a Poisson regression of each name generator was performed on household size, sex, education, age and mode of administration. Note that only those respondents are included in the analysis who answered *"yes"* to the filter question *"Does anybody ELSE live with you in this household? "*. A number of these respondents named no alters even though they answered *"yes"* to the filter question. Descriptive information is presented in Table 31 and the results of the regression analysis in Table 32. No significant mode effects on the number of alters mentioned is visible.

Table 31 Row percentages and descriptives for items a301b, a302b, and for both combined

	0	1	2	3	4	5	6+	n	mean
a301 F2F	7.1	25.0	39.3	19.6	8.9	0.0	0.0	56	1.98
a301 CATI	0.0	34.2	35.6	20.5	4.1	1.4	4.1	73	2.19
a301 WEB	0.0	44.9	18.8	15.9	14.5	2.9	2.9	69	2.23
a302 F2F	7.1	50.0	35.7	7.1	0.0	0.0	0.0	14	1.43
a302 CATI	0.0	85.7	14.3	0.0	0.0	0.0	0.0	7	1.14
a302 WEB	22.7	36.4	4.5	13.6	4.5	9.1	9.1	22	2.14
all F2F	5.0	28.3	40.0	18.3	6.7	1.7	0.0	60	1.98
all CATI	0.0	36.0	34.7	20.0	4.0	1.3	4.0	75	2.16
all WEB	5.4	43.2	16.2	16.2	12.2	2.7	4.1	74	2.19

Table 32 Poisson regression for size of »other household members« network

Variable	b	s.e	t	p
a301b: *other household members (currently not away)*				
CATI	0.18	0.13	1.41	0.16
WEB	0.12	0.13	0.96	0.34
a302b: *other household members (currently away)*				
CATI	-0.27	0.44	-0.61	0.54
WEB	0.24	0.29	0.82	0.41
a301b+a302b				
CATI	0.16	0.12	1.26	0.21
WEB	0.11	0.12	0.85	0.39

We now move to the analysis of the name interpreters. Item non response is denoted by the acronym INR in the tables. In each table, two types of percentages are given:

- A-type percentages refer to the proportion of alters that fall within a certain category. E.g. in Table 6, 55.1% of alters (across all modes) were biological or adoptive parents. In Total, 425 alters were named by all respondents combined. The problem with regarding the alter as the unit of analysis is that respondents who named more alters have a larger bearing on the results.
- B-type percentages aim to correct this and refer to the number of all respondents who named at least one alter that falls into the given category. E.g. in Table 6, 72.7% of respondents (out of 198 respondents who received the question) named at least one parent. Note that B-type percentages need not sum up to 100.

Table 33 shows the results for the following chi squared tests:

- sex of alters named at a301b by mode,
- categorized age of alters named at a301b by mode,
- sex of alters named at a302b by mode,
- categorized age of alters named at a302b by mode.

Put differently, the chi squared tests were performed on tables of A-type frequencies that correspond to Tables 7, 8, 11, and 12. The INR (item nonresponse) row was not part of the test. None of the four tests show significant mode differences in alters' sex and age composition.

Table 33 Chi squared test for sex and age composition of alters by mode

	chi square	DF	sig.	Cramer's V
a301b sex	1.60	2	0.45	0.06
a301b age	17.71	12	0.12	0.14
a302b sex	1.52	2	0.47	0.14
a302b age	10.75	12	0.55	0.27

Results of both regression analysis and chi-square test provide enough evidence to claim that there are no significant mode effect for the egocentric networks, pertaining to household composition.

5.2.2 Network delineation and support

The following name generators were analyzed:

a503: taking care of children within the household
a505: receiving childcare help
a511: provision of childcare help
a513: taking care of the practicalities around the house
a515: receiving help with household tasks
a519: provision of help with household tasks
a528: receiving personal care
a530: provision of personal care
a532: provision of financial support
a535: received a contribution or inherited money worth more than 500€

Most respondents in web mode followed the instructions and typed in the first name and initial of the last name for each alter. Ten respondents misunderstood the instructions and

entered *several* alters' names at once (e.g. "Jane and John D."). Five respondents did not want to give the name and initial and instead entered two initials, only one letter etc. Less than five respondents did not follow the instructions and typed in the relation to the alter (e.g. "sister", "mother") instead of the name and initial. For item a535 (received a contribution or inherited money, goods, or property worth more than 5000€) one common answer was "guests at our wedding".

In those cases where the "name" of the alter actually refers to several people, the subsequent name interpreter items are meaningless e.g. "Please indicate whether the person *guests at our wedding* is male or female".

Results on the analysis of mode effects on the size of each network were presented in the section (4.4) on the methodological experiments. In this section, we present the results of network composition by comparing the structural percentages. It should be noted that items a503 (taking care of children within the household) and a513 (taking care of practicalities around the house) included the option "Me" - the respondent could nominate himself/herself. Such answers are *not* counted as alters in the analyses in this section. Tables 34 and 35 give the frequencies of respondents who included "Me" as an answer to items a503 and a513.

Table 34 Proportion of respondents who included the answer 'Me' at item a503 (taking care of children within the household)

	F2F	CATI	WEB	Total
no 'Me' answer	25.0	58.3	15.8	30.8
includes 'Me'	75.0	41.7	84.2	69.2
Total n	8	12	19	39

Table 35 Proportion of respondents who included the answer 'Me' at item a513 (taking are of practicalities within the household)

	F2F	CATI	WEB	Total
no 'Me' answer	41.7	30.8	21.7	31.0
includes 'Me'	58.3	69.2	78.3	69.0
Total n	144	146	161	451

In both cases, more than half the respondents (also) nominated themselves.

Table 36 gives the frequencies and descriptives for each name generator by mode.

Table 36 Frequencies and descriptives for name generators in Module on social networks

	0	1	2	3	4	5	6+	n	mean
a501 F2F	4.4	36.4	32.0	12.6	8.3	3.9	2.4	206	2.09
a501 CATI	7.5	21.0	19.5	23.0	15.0	11.0	3.0	200	2.66
a501 WEB	7.0	37.7	17.2	12.1	12.1	10.7	3.3	215	2.32
a503 F2F	37.5	37.5	25.0	0.0	0.0	0.0	0.0	8	0.88
a503 CATI	16.7	33.3	33.3	8.3	8.3	0.0	0.0	12	1.58
a503 WEB	10.5	31.6	31.6	15.8	0.0	5.3	5.3	19	2.00
a505 F2F	4.0	24.0	44.0	8.0	16.0	0.0	4.0	25	2.28
a505 CATI	0.0	25.8	35.5	22.6	12.9	0.0	3.2	31	2.35
a505 WEB	2.3	39.5	25.6	18.6	7.0	4.7	2.3	43	2.19
a511 F2F	72.3	19.4	5.3	2.4	0.5	0.0	0.0	206	0.39
a511 CATI	64.0	20.5	12.5	2.5	0.5	0.0	0.0	200	0.55
a511 WEB	68.4	24.2	6.5	0.5	0.0	0.5	0.0	215	0.41
a513 F2F	36.8	47.2	13.2	2.8	0.0	0.0	0.0	144	0.82
a513 CATI	28.1	45.9	17.8	6.8	1.4	0.0	0.0	146	1.08
a513 WEB	33.5	46.6	15.5	3.1	0.6	0.6	0.0	161	0.93
a515 F2F	90.3	6.3	2.9	0.5	0.0	0.0	0.0	206	0.14
a515 CATI	91.0	6.5	1.5	0.5	0.0	0.5	0.0	200	0.14
a515 WEB	82.3	13.0	3.3	0.5	0.9	0.0	0.0	215	0.25
a519 F2F	87.4	11.2	1.5	0.0	0.0	0.0	0.0	206	0.14
a519 CATI	81.5	10.5	4.5	3.0	0.5	0.0	0.0	200	0.30
a519 WEB	83.7	12.6	2.3	0.5	0.5	0.5	0.0	215	0.23
a528 F2F	95.1	3.9	1.0	0.0	0.0	0.0	0.0	206	0.06
a528 CATI	88.0	9.0	2.0	1.0	0.0	0.0	0.0	200	0.16
a528 WEB	88.8	8.4	1.4	1.4	0.0	0.0	0.0	215	0.15
a530 F2F	90.3	6.8	2.4	0.5	0.0	0.0	0.0	206	0.13
a530 CATI	91.5	4.0	3.0	0.5	0.5	0.0	0.5	200	0.16
a530 WEB	88.4	7.0	3.3	0.5	0.9	0.0	0.0	215	0.19
a532 F2F	90.8	6.3	2.4	0.5	0.0	0.0	0.0	206	0.13
a532 CATI	88.5	7.0	3.5	1.0	0.0	0.0	0.0	200	0.17
a532 WEB	94.4	4.7	0.5	0.5	0.0	0.0	0.0	215	0.07
a535 F2F	77.7	19.4	2.9	0.0	0.0	0.0	0.0	206	0.25

a535 CATI	76.0	21.0	3.0	0.0	0.0	0.0	0.0	200	0.27
a535 WEB	80.0	17.2	2.8	0.0	0.0	0.0	0.0	215	0.23

As we already found out in the section on methodological experiments, some mode effects on the number of alters named exists. In most cases, the respondents named *more* alters in CATI and web modes compared to face-to-face (the only exception is item a532 where web administration resulted in fewer alters named).

The newly proposed measures for egocentric networks, based on the name generator and name interpreter approach behave well in the pilot data, thus the recommendation is to use this sort of scales for the GGS questionnaire.

5.3 Health and Well-being scales

Potential scales were analyzed by investigating metric and scalar equivalence across modes. Scalar equivalence is needed for conclusions based on differences between subgroups across modes. More specifically it pertains to the equivalence of means of scales across three different modes. Metric equivalence on the other hand pertains to the invariance of factor loadings of individual items of a scale across different modes (Van de Schoot et al., 2012). The strategy was to start with the strictest equivalence (both metric and scalar), which was in case of low fit relaxed.

5.3.1 Locus of control

The following item set was analyzed:

a706a: There is really no way I can solve some of the problems I have
a706b: Sometimes I feel that I'm being pushed around in life
a706c: I have little control over the things that happen to me
a706d: I often feel helpless in dealing with the problems of life
a706e: There is little I can do to change many of the important things in my life

Table 37 Fit indices for metric and scalar equivalence across modes for the locus of control scale

	Chi-Square	df	*p*	CFI	TLI	RMSEA
Metric	59.379	23	<.0001	0.956	0.942	0.088
Scalar	88.176	31	<.0001	0.930	0.932	0.095
Scalar *without item a*	54.734	18	<.0001	0.942	0.942	0.099

The results in Table 37 suggest that all fit indices are acceptable, but suggest that the scale could be improved. The scale fits a bit better without item a706a, but the improvement is not significant.

Table 38 Reliabilities of the locus of control scale

	Face to Face	Telephone	Web
N	203	200	208
Chronbach's alpha	.688	.741	.833
Mean correlation	.360	.418	.556

The locus of control scale looks fine both for equivalence and reliability (for the latter see Table 38). If item a706a is omitted fit and reliability remain about the same. **We recommend to retain this scale and consider excluding item a.**

5.3.2 Loneliness

The following item set was analyzed:

a708a: There are plenty of people that I can lean on in case of trouble
a708b: I experience a general sense of emptiness
a708c: I miss having people around
a708d: There are many people that I can count on completely
a708e: Often, I feel rejected
a708f: There are enough people that I feel close to

A one factor confirmatory factor analysis with all variables labeled as categorical, results in non-positive definite warning for F2F mode. A two factor model also resulted in non-positive definite warnings for mode face to face and telephone. When item d is removed, the fit is better. Reliabilities are not high, but appear acceptable (see Table 39). We suspect there may be a genuine mode effect here.

Table 39 Reliabilities of the loneliness scale

	Face to Face	Telephone	Web
N	204	200	213
Chronbach's alpha	.666	.685	.714
Mean correlation	.252	.273	.295

Our recommendation is to retain the scale, including item d. In data collection, the respondents in the F2F condition should provide their answers in self completion mode.

5.3.3. Depression

The following item set was analyzed:

a709a: I felt that I could not shake off the blues even with help from my family or friends
a709b: I felt depressed
a709c: I thought my life had been a failure
a709d: I felt fearful
a709e: I felt lonely
a709f: I had crying spells
a709g: I felt sad

Table 40 Fit indices for metric and scalar equivalence across modes for the depression scale

	Chi-Square	Df	*p*	CFI	TLI	RMSEA
Metric	215.661	54	<.0001	0.901	0.884	0.120
Scalar	245.956	66	<.0001	0.889	0.894	0.115
Partial Scalar*	212.704	62	<.0001	0.907	0.906	0.108

*released factor loadings per group for items c, f, g.

Results in Table 40 suggest that the one factor model is on the margin of being acceptable. A multigroup confirmatory factor analysis with one factor was run to examine the scale of depression. There is only weak support of both metric and scalar invariance. The partial scalar model, which releases equivalence constraints of the factor loadings on some of the items (a,f,g) performs best. However, the results of the reliability analyses presented in Table 41 show that Cronbach's alpha for all of the items together, is acceptable.

Table 41 Reliabilities of the depression scale

	Face to Face	Telephone	Web
N	204	198	214
Chronbach's alpha	.838	.827	.882
Mean correlation	.430	.410	.516

For this scale, the factor means of the three modes were significantly different. We suspect there may be a genuine mode effect here. If items g and f are removed, the fit becomes better. **Our recommendation is to retain the scale, and recommend users to use SEM analysis instead of working with sum scores. Consider dropping items e and f. In data collection, let respondents in face-to-face condition provide their answers in self completion mode.**

5.4 Personality scales

A 15-item short version of the Big Five Inventory was used (Rammstedt and John, 2007). The following dimensions and items were included in the analysis:

Extraversion: A705c, a705h, a705m(R)
Agreeableness: a705a(R), a705f, a705k
Conscientiousness: a705b, a705g(R), a705l
Neuroticism: a705d, a705i, a705n(R)
Openness: a705e, a705j, a705o

Results from confirmatory factor and reliability analyses on these five dimensions are presented in Table 42.

Table 42 Fit indices of CFA for »big five« scales and their reliabilities

	Chi-square	df	p	CFI	TLI	RMSEA	α(F2F)	α(CATI)	α(WEB)
Extraversion	18.9	8	0.01	0.97	0.96	0.08	0.72	0.63	0.71
Agreeableness	12.1	8	0.14	0.98	0.97	0.05	0.50	0.50	0.58
Conscientiousness*							0.29	0.32	0.30
Neuroticism	18.9	8	0.01	0.95	0.94	0.08	0.68	0.40	0.68
Openness	26.6	8	0.01	0.93	0.92	0.11	0.60	0.60	0.68

*no convergence for metric or scalar equivalence, using different estimators

The results show that for *Extraversion* the measurement equivalence and reliabilities are satisfactory. **The recommendation is to retain the scale.** For *Agreeableness* measurement equivalence demonstrates good fit, but reliabilities are rather low. Nevertheless, **the recommendation is to retain the scale, but users should rather use SEM analysis instead of working with sum scores.** *Conscientiousness* seems to be problematic, as CFA does not converge and the reliabilities are very low. **The recommendation would be to replace item a705g, possibly also item a705b.** Consult Rammstedt and John (2007) for candidate replacement items. *Neuroticism* demonstrates satisfactory measurement equivalence, while reliability for CATI mode is low. **Recommendation would be to retain the scale, and recommend users to use SEM analysis instead of working with sum scores.** The measurement equivalence and reliabilities for *Openness* are acceptable, but Cronbach's alpha is not very high. **Our recommendation is to retain the scale, and recommend users to use SEM analysis instead of working with sum scores.**

5.5 Value orientations and attitudes

5.5.1 Religious symbolism

The following items were analyzed:

a11.03a: It is important for an infant to be registered in the appropriate religious ceremony
a11.03b: It is important for people who marry in registry offices to have a religious wedding too
a11.03c: It is important for a funeral to include a religious ceremony

Results from confirmatory factor analyses are presented in Table 43, and from reliability analyses in Table 44. Scalar measurement invariance holds across the three modes of data collection for this scale. Reliability is fine. **Our recommendation is to retain the scale.**

Table 43 Fit indices for metric and scalar equivalence across modes for the religious symbolism scale

	Chi-Square	df	p	CFI	TLI	RMSEA
Scalar	15.182	10	0.1256	0.995	0.996	0.050
Full	61.104	16	<.0001	0.959	0.977	0.117

Table 44 Reliabilities of the traditionalism scale

	Face to Face	Telephone	Web
N	206	200	213
Chronbach's alpha	.845	.880	.925
Mean correlation	.652	.715	.805

5.5.2 Family values/traditionalism

This scale is composed of the following attitude items:

a1108a: Marriage is an outdated institution
a1108b: It is all right for an unmarried couple to live together even if they have no interest in marriage
a1108c: Marriage is a lifetime relationship and should never be ended
a1108d: It is all right for a couple with an unhappy marriage to get a divorce even if they have children
a1108e: A woman has to have children in order to be fulfilled
a1108f: A man has to have children in order to be fulfilled
a1108g: A child needs a home with both a father and a mother to grow up happily
a1108h: A woman can have a child as a single parent even if she doesn't want to have a stable relationship with a man
a1108i: Homosexual couples should have the same rights as heterosexual couples do

Table 45 Fit indices for metric and scalar equivalence across modes for the traditionalism scale

	Chi-Square	df	p	CFI	TLI	RMSEA
Metric	181.237	74	<.0001	0.869	0.852	0.084
Metric partial*	205.117	88	<.0001	0.867	0.864	0.080
Scalar	221.368	90	<.0001	0.840	0.851	0.084
Scalar 2F	179.678	81	<.0001	0.880	0.875	0.077

*released intercept for item h, based on modification index.

A two-factor CFA was conducted based on the EFA loadings, with items b, d, h as a first factor and items a, c, e, g, and i, as a second factor. Results are presented in Table 45. The multicollinear item (f) was excluded from this analysis. Scalar measurement equivalence holds (excluding one multi-collinear item).

Table 46 Reliabilities of the traditionalism scale

	Face to Face	Telephone	Web
N	204	198	214
Chronbach's alpha	.701	.752	.816
Mean correlation	.210	.253	.336

Reliabilities are sufficient (see Table 46). As expected, reliabilities are somewhat higher when the multicollinear item is included. A two-factor model did not fit much better. **Our recommendation is to retain the scale and drop item e because of redundancy (multicollinearity with item a1108f).**

5.5.3 Institutional support arrangements

This scale consists of the following set of items:

a1109a: care for older persons in need of care at their home
a1109b: care for pre-school children
a1109c: care for school children during after-school hours
a1109d: financial support for older people below subsistence level
a1109e: financial support for younger people below subsistence level

In GGS wave 1, a two factor solution was found to be appropriate for this scale. There is no multicollinearity. In order to examine measurement equivalence across modes, a CFA with two correlated factors was conducted. Results are presented in Table 47. Based on the above results, measurement invariance holds across the three modes for this scale, including equivalent correlations between factors across the three groups.

Table 47 Fit indices for metric and scalar equivalence across modes for the institutional support

	Chi-Square	df	p	CFI	TLI	RMSEA
Metric	45.864	18	0.0003	0.947	0.911	0.086
Scalar	50.765	24	0.0011	0.949	0.936	0.073
Scalar-2*	51.195	26	0.0023	0.952	0.944	0.068

* Scalar equivalence, but also having the correlations between the two factors equal across the three modes

Table 48 Reliabilities of the institutional support scale

	Face to Face	Telephone	Web
N	206	199	213
Chronbach's alpha	.585	.506	.686
Mean correlation	.214	.170	.302

Cronbach's alphas, as shown in Table 48 are rather low, but highest for the web mode. **Our recommendation is to retain the scale, and recommend users to use SEM analysis instead of working with sum scores.**

5.5.4 Parental obligations

This scale refers to items a1110a to a1110c. These items are:

a1110a: grandparents should look after their grandchildren if the parents of these grandchildren are unable to do so

a1110b: parents ought to provide financial help for their adult children when the children are having financial difficulties
a1110c: if their adult children were in need, parents should adjust their own lives in order to help them

Table 49 Fit indices for metric and scalar equivalence across modes for the parental obligations

	Chi-Square	df	p	CFI	TLI	RMSEA
Scalar	5.564	8	0.7017	1.000	1.006	0.000
Full	58.211	14	<.0001	0.909	0.941	0.124

Table 50 Reliabilities of the parental obligations scale

	Face to Face	Telephone	Web
N	205	200	213
Chronbach's alpha	.767	.662	.807
Mean correlation	.533	.403	.585

Results in Table 49 show that scalar measurement equivalence holds across modes for this scale. Results in Table 50 show that reliabilities are acceptable. **Our recommendation is to retain the scale.**

5.5.5 Filial obligations

This scale includes items a1111a to a1111d:

a1111a: children should take responsibility for caring for their parents when parents are in need
a1111b: children should adjust their working lives to the needs of their parents
a1111c: children ought to provide financial help for their parents when their parents are having financial difficulties
a1111d: children should have their parents to live with them when parents can no longer look after themselves

Results of the confirmatory factor analyses are presented in Table 51. The fit does not look good. There is no measurement equivalence across modes. According to the modification indices, the intercepts of items a and c were problematic. The model with different intercepts for items a and c, appears best out of the three, but still on the margin of being acceptable. Results in table 52 suggest that the reliability of the scale in the different modes is acceptable. **Our recommendation is to consider replacing this scale. If this is not feasible, retain the scale, but warn users that analyses involving this scale may be hindered by mode issues.**

Table 51 Fit indices for metric and scalar equivalence across modes for the filial obligations

	Chi-Square	df	p	CFI	TLI	RMSEA
Metric	77.272	12	<.0001	0.844	0.767	0.162
Partial Metric*	77.903	14	<.0001	0.848	0.804	0.149
Scalar	101.840	18	<.0001	0.800	0.800	0.150

*intercepts for items a and c released across mode

Table 52 Reliabilities of the filial obligations scale

	Face to Face	Telephone	Web
N	206	198	214
Chronbach's alpha	.627	.607	.746
Mean correlation	.296	.277	.429

5.5.6 Gender attitudes

This scale refers to items a1112a to a1112h:

a1112a: Work is good, but what most women really want is a home and children
a1112b: Being a housewife is just as fulfilling as working
a1112c: It is the task of a man to earn money and that of a woman to look after the home and the family
a1112d: It is not good if the man stays at home and cares for the children and woman goes out to work

a1112e: The relationship between a working woman and her children can be just as close as that of a non-working mother
a1112f: A pre-school child will probably suffer if his/her mother works
a1112g: All in all family life suffers if the woman works full-time
a1112h: Family life often suffers because men concentrate too much on their work

Table 53 Fit indices for metric and scalar equivalence across modes for the gender attitudes

	Chi-Square	df	p	CFI	TLI	RMSEA
Metric	157.172	69	<.0001	0.912	0.893	0.079
Scalar	236.325	81	<.0001	0.846	0.840	0.096

Based on the results presented in Table 53, fit looks best for two-factor model. This consists of items a, b, and d (factor: housework) and of the remainder (factor: consequences for family). A CFA model for two factors indicates sufficient measurement equivalence, but rather low reliabilities for the first factor (which has only 3 items). For this scale, the factor means of the three modes were significantly different (but less so than for depression). We suspect there may be a genuine mode effect here.

Table 54 Reliabilities of the gender attitudes scale

	Face to Face	Telephone	Web
N	204	199	213
Chronbach's alpha	.597	.710	.755
Mean correlation	.168	.241	.292

Based on the results presented in Table 54, reliability looks acceptable. **Our recommendation is to retain the scale, and suggest users to use SEM analysis instead of working with sum scores. In data collection, let respondents in face-to-face condition provide their answers in self completion mode.**

6. Theory of planned behavior items

The items that pertain to intentions of conducting various activities were already analyzed in the context of the methodological experiments reported on in section 4.2. In this section, we focus on the more specific TPB concepts of *attitudes* toward a certain activity, *perceived behavior control* over activity and *subjective norm* about the activity. On all scales we performed CFA and reliability analyses. The results are presented in condensed form in Table 55.

Table 55 Fit statistics of the measurement equivalence and reliabilities for the theory of planned behavior scales

Concept	items	fit	alpha
Entry into a union			
Attitude about entry into a union	a280a to a280d	$\chi^2(2) = 0.872$, p= 0.6467	0.63
Perceived behavioral control over entry into a union	a281a to a281d	$\chi^2(2) = 6.197$, p= 0.0451	0.76
Subjective norm about entry into a union	a282a to a282d	$\chi^2(2) = 0.$ 0.174, p= 0.9169	0.84
Leaving the parental home			
Attitude about leaving the parental home	a449a to a449e	$\chi^2(2) = 0.058$, p= 0.9717	0.76
Perceived behavioral control over the parental home	a450a to a450d	$\chi^2() = 0.524$, p= 0.7695	0.79
Subjective norm about leaving the parental home	a451a to a451d	$\chi^2() = 1.184$, p= 0.5532	0.84
Having Children			
Attitude about having a/another child	a619a to a619e	$\chi^2(2) = 8.872$, p= 0.0118	0.74
Perceived behavioral control over having a/another child	a620a to a620i	2 factor solution: $\chi^2(4) =$ 14.282, p= 0.0064	0.89
Subjective norm about having a/another child	a621a to a621d	$\chi^2(2) = 3.260$, p= 0.1959	0.90
Employment and Retirement			
Attitudes about retirement	a835a to a835e	$\chi^2(2) = 9.717$, p= 0.0078	0.82

Perceived behavioral control over retirement	a836a to a836d	$\chi^2(2) = 0.600$, p= 0.7409	0.85
Subjective norm about retirement	a837a to a837f		0.90

It should first be noted that for most behavioral domains, the number of cases was small, which makes detailed analyses difficult. Despite that, it can be claimed that in all cases, the measurement equivalence and reliabilities of the attitude scales look good. For subjective norms and perceived behavioral control it was not always clear if these formed a scale or should be viewed as an index.

Some specific observations and recommendations:

Entry into union:

- Attitudes: Retain all items, but note that attitude on financial situation does not sit well with the other items. The wording might not convey to all respondents the concept we are interested in;
- Perceived behavior control: Retain all items, but consider for future surveys whether this item set could be improved as a scale by (i) expressing affordability and housing in terms of a common concept, and (ii) reviewing the role of readiness, which dominated the measurement model;
- Subjective norm: Retain all items, but consider revising the subjective norm item that pertains to "most of my friends" (a282a) to "my closest friends" to better convey this concept and to distinguish it more clearly from other items in the scale.

Leaving home:

- Attitudes: Retain all items, but note that the attitude about the financial situation does not sit well with the other items, and cross loads on both subjective norms and perceived behavioral control. It is not a good item to include in attitudes scales;
- Perceived behavioral control: Retain all items, but note that the several issues raised in the previous summary of PBC for items on entry into a union, and flag this

item bank for further attention in later revisions. Consider, in particular, the role of a450d;

- Subjective norm: Retain all items, but note that several imperfections in the formation of scales were observed, and in particular that item a451d is multicollinear, and the error variance needs to be constrained in order to reach a CFA solution.

Having a child:

- Attitudes: Retain all items, but note the low variation of item a619b on the 'financial situation', as well as its low loading on attitude for males;
- Perceived behavioral control: Add back item a620g on partner's health. Flag the high error variance between items a620h and a620i for monitoring in a next review ;
- Subjective norm: Retain all items, but note that several imperfections in the formation of scales were observed, and in particular that item a621d is highly correlated with other subjective norms items

Retirement:

- Attitudes: Retain all items, but note that the attitude toward financial situation does not sit well with the other items. It is not a good item to include in Attitudes scales;
- Perceived behavioral control: Retain all items, but simplify the wording of item a836c by omitting the redundant ending 'from the workforce';
- Subjective norm: Retain all items.

REFERENCES

Bollen, K.A. (1989), *Structural Equation Models with Latent Variables*. Wiley & Sons, New York.

Brown, T.A. (2006), *Confirmatory Factor Analysis for Applied Research.*New York: Guilford press.

Browne, M.W. and R. Cudeck (1993), Alternative ways of assessing model fit. In: Bollen KA, Long JS, (eds.): *Testing Structural Equation Models.* Beverly Hills, CA: Sage, pp. 136-162.

Cameron, A.C. and P.K. Trivedi (1998), *Regression analysis of count data,* Cambridge University Press.

Crawford, S.D., M.P. Couper and M.J. Lamias (2001), Web surveys: Perception of Burden. *Social Science Computer Review*, 19(2): 146-162.

DeVellis, R.F. (2012), *Scale development: Theory and Applications*. London, UK: Sage.

Hair, J.F., R.E. Anderson, R.L. Tatham and W.C. Black (1998), *Multivariate Data Analysis*, Prentice Hall of India Private Limited, New Delhi.

Hu, L. and P.M. Bentler (1999), Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6:1-55.

Hox, J.J. (2010), Multilevel analysis. Techniques and applications. New York: Routledge. http://www.taylorandfrancis.com/books/details/9781848728462/

Kogovšsek, T. and A. Ferligoj (2005), Effects on reliability and validity of egocentered network measurements. *Social Networks* 27(3), 205-229.

Schuman, H. and S. Presser (1981), *Questions and Answers in Attitude Surveys: Experiments on Question Form, Order, and Context*. New York: Academic Press.

Sharp, L.M. and J. Frankel (1983), Respondent burden: A test of some common assumptions. *Public Opinion quarterly*, 47(1): 36-53.

Van de Vijver, F. (1998), Towards a Theory of Bias and Equivalence. *Zuma Nachrichten: Cross-Cultural Survey Equivalence*, 3, 41-65.

Van de Schoot, R., Lugtig, P. and J.J. Hox (2012), A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-292.