DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 34, ARTICLE 18, PAGES 499–524
PUBLISHED 15 MARCH 2016
http://www.demographic-research.org/Volumes/Vol34/18/
DOI: 10.4054/DemRes.2016.34.18

Research Article

Generations and Gender Programme Wave 1 data collection: An overview and assessment of sampling and fieldwork methods, weighting procedures, and cross-sectional representativeness

Tineke Fokkema **Tom Emery**

Andrej Kveder **Aart C. Liefbroer**

Nicole Hiekel

This publication is part of the Special Collection on "Data Quality Issues in the First Wave of the Generations and Gender Survey," organized by Guest Editors Aart C. Liefbroer and Joop Hox.

© 2016 Tineke Fokkema et al.

This open-access work is published under the terms of the Creative Commons Attribution NonCommercial License 2.0 Germany, which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit.
See http://creativecommons.org/licenses/by-nc/2.0/de/

Demographic Research: Volume 34, Article 18
Research Article

Generations and Gender Programme Wave 1 data collection: An overview and assessment of sampling and fieldwork methods, weighting procedures, and cross-sectional representativeness

Tineke Fokkema[1,2,3]

Andrej Kveder[4]

Nicole Hiekel[5]

Tom Emery[1,2,3]

Aart C. Liefbroer[1,6,7]

Abstract

BACKGROUND
The Generations and Gender Survey (GGS) was developed to stimulate the study of a broad range of topics of relevance to population scientists. So far, at least one wave of the GGS has been conducted in 19 countries. If scholars want to use the GGS for comparative purposes, it is essential that there be cross-national equivalence in terms of survey implementation and representativeness.

OBJECTIVES
The two main goals are (1) to describe the main features of the implementation of the GGS in participating countries, and (2) to describe and evaluate the quality of the data collection of the GGS in terms of its cross-sectional representativeness.

METHODS
We use weighted and unweighted GGS data for 18 countries and compare this to country-specific information.

[1] Netherlands Interdisciplinary Demographic Institute, The Hague, the Netherlands.
E-Mail: fokkema@nidi.nl.
[2] University of Groningen, the Netherlands.
[3] Department of Sociology, Erasmus University Rotterdam, the Netherlands.
[4] Oxford Policy Management, Oxford, United Kingdom.
[5] Institute of Sociology and Social Psychology, University of Cologne, Germany.
[6] University Medical Center Groningen, Groningen, the Netherlands.
[7] Department of Sociology, Vrije Universiteit, Amsterdam, the Netherlands.

RESULTS
The quality of sampling and fieldwork procedures of the GGS is generally good. On average, response rates in the GGS are comparable to those in other cross-national surveys. After weighting, the data are generally representative in terms of age, gender, region, and household size, but less so for marital status and educational attainment. Implications for future waves of the GGS are discussed.

1. Introduction

The Generations and Gender Survey (GGS) was developed to stimulate the study of a broad range of topics of relevance to population scientists. It is designed as a three-wave panel study conducted at three-year time intervals across as many developed countries as possible, and covers a range of demographically relevant topics such as leaving home, union formation, fertility decision-making, combining employment and parenthood, intergenerational solidarity, and retirement. So far, at least one wave of the GGS has been conducted in 19 countries.

If scholars want to use the GGS for comparative purposes it is not just essential that the quality of the data collected in each specific country be high, but also that there be cross-national equivalence in terms of survey implementation and representativeness. The Generations and Gender Programme (GGP) provides a broad set of country-specific documentation to facilitate the understanding of data collection procedures and data quality for each participating country.[8] However, a comparative description and analysis of the quality of data collection is still lacking. Against this backdrop, this article has two main goals:

1. To describe the main features of the implementation of the GGS in participating countries; and
2. To describe and evaluate the quality of GGS data collection in terms of its cross-sectional representativeness.

To achieve these goals, we will first describe and discuss characteristics of the sampling procedures. Next, fieldwork procedures and response rates achieved will be discussed. Finally, attention will be paid to the cross-sectional representativeness of the datasets by reviewing the weighting procedures applied to correct for design effects and for post-stratification differences between the national samples and national

[8] See information on the GGS Data Description pages of the GGP website (http://www.ggp-i.org/materials/ggs-data-description.html).

populations. We will also compare explicitly – both before and after weighting – the distribution of our national samples on a set of key characteristics with population estimates based on national census data and other official sources. The article will conclude with a summary of the main findings and recommendations for data users.

2. Sample

Clear prescriptions about the main sampling characteristics have been developed within the Generations and Gender Programme (GGP) by Simard and Franklin (2005). The three most important elements of these sample design guidelines were:

1. The target population in a country is the resident non-institutionalised population aged 18–79 at the time of the first wave;
2. The sample size of wave 1 should be high enough to interview at least 8,000 respondents in wave 3. In general, a realised sample size of at least 10,000 in wave 1 was deemed necessary;
3. Apply probability sampling. The exact method used was allowed to vary across countries based on the availability and cost-effectiveness of different sampling frames.

Table 1 presents information on sampling for wave 1 country datasets that have thus far been released. The first two columns give information on whether one-stage or multi-stage sampling was applied and whether stratification was applied. A one-stage procedure was used in only five countries (Austria, Estonia, the Netherlands, Norway, and Sweden), with respondents being drawn without first selecting higher-order units. In all other countries except Australia a two-stage sampling strategy was used. In a first stage, areas were selected followed by a selection of individual sample elements – names, addresses, or dwellings. In Australia a three-stage procedure was used: dwellings were selected within selected areas, followed by a random sample of three households if a dwelling was occupied by four or more households. Stratification was applied in the majority of countries. Only in Bulgaria, Germany, the Netherlands, Romania, and the Russian Federation was no stratification used.

The next two columns present information on the sampling frame and the frame elements. Basically, three types of approach can be distinguished. In a first group of countries (Austria, Belgium, Italy, Norway, and Sweden), population registers were used as the sampling frame and names constituted the frame elements. Clearly, this approach is only feasible in countries where population registers exist and are accessible to social science research. In a second group of countries (Australia, the

Czech Republic, Germany, Hungary, Lithuania, the Netherlands, Romania, and the Russian Federation) area sampling was used with either addresses or dwellings as sampling elements. Finally, in a third group of countries (Bulgaria, Estonia, France, Georgia, and Poland), (a combination of area and) census information was used as the sampling frame, with either names or dwellings as sampling elements.

Table 1: Main sampling characteristics of wave 1 of the Generations and Gender Survey, by participating country

Country	# sampling stages	Stratification	Frame	Frame elements	Type of sampling	Sample size
Australia	3	YES	Area	Dwellings, Households	PPS + SRS	13,571
Austria	1	YES	Population register	Names	SRS	9,006
Belgium	2	YES	Population register	Names	SRS	17,836
Bulgaria	2	NO	Area & Census	Dwellings	PPS + SRS	18,591
Czech Republic	2	YES	Area	Dwellings	PPS + SRS	23,824
Estonia	1	YES	Census	Names	SRS	11,192
France	2	YES	Census & update new dwellings	Dwellings	PPS + SRS	18,009
Georgia	2	YES	Census	Names	PPS + SRS	14,000
Germany	2	NO	Area (GIS)	Addresses	PPS + SRS	20,623
Hungary	2	YES	Area, Settlement	Addresses	PPS + SRS	24,138
Italy	2	YES	Population register	Names	PPS + SysR	20,787
Lithuania	2	YES	Area	Settlements	RR + RR	29,884
Netherlands	1	NO	Area	Addresses	SRS	24,434
Norway	1	YES	Population register	Names	SRS	25,848
Poland	2	YES	Census area	Dwellings	SRS	60,000
Romania	2	NO	Area	Dwellings	SRS	14,280
Russian Federation	2	NO	Area	Dwellings	PPS + SRS	27,089
Sweden	1	YES	Population register	Names	SRS	18,000
Germany-Turks	2	NO	Local and federal register of foreigners	Addresses	PPS + SRS	13,890

PPS = Probability Proportional to Size, SRS = Simple Random Sampling, RR = Random Route, SysR = Systematic sampling with a Random start

The fifth column shows that a probability sampling procedure was applied in all countries. Simple Random Sampling (SRS) procedures were applied in all five countries that used one-stage sampling: Austria, Estonia, the Netherlands, Norway, and Sweden. In most countries that applied a two-stage sampling procedure the sampling of higher-order units was done using Probability Proportional to Size (PPS), where the chance of a higher-order unit (here: municipalities or other regional units) being selected is related to the relative proportion of lower-order units (here: persons) in the higher-order unit. Within the selected higher-order unit, SRS was used to select individual sample members. [9] Random Route (RR) methods were used only in

[9] In Italy, systematic sampling with a random start (SysR) was used rather than SRS.

Lithuania. The final column shows the size of the total sample used for data collection. It varied from 9,006 persons in Austria to 60,000 in Poland.

The overall conclusion to be drawn from Table 1 is that all countries applied the prescription of drawing a probability sample. The exact methods by which this was done depended on country-specific conditions. In those countries where it was feasible to draw a simple random sample from national registers or census information, this was the preferred and most frequently used method. However, in some countries this was not feasible, and other randomised methods had to be employed.

3. Fieldwork

The comparability of international surveys also depends on similarity in the fieldwork procedures used. No stringent prescriptions for these procedures were developed for the GGS. Rather, a set of fieldwork guidelines were issued that were "meant as a collection of good practices" to support the survey fieldwork (Kveder 2007, p. 47). These guidelines mainly included advice on interview training and contacting procedures. Recommendations to optimise panel maintenance were also made. The non-binding nature of these guidelines, together with the fact that they only became available after some countries had already conducted their first wave of data collection and that what is deemed good practice may vary by country and fieldwork agency, stresses the importance of achieving a good overview of how the fieldwork was organised in participating countries. Information on the most important aspects are summarised in Tables 2 and 3.

Table 2 provides information on population coverage and data collection. As mentioned in section 2, the sampling guidelines recommended focusing on the non-institutionalised population aged 18 to 79. Most of the participating countries stuck to the suggested age guideline. Small deviations were observed in Estonia and Hungary, where the lower age range of the samples was 21 and the upper age range for the Estonian sample was 80. In Australia all respondents aged 15 and older were covered. The largest deviations in the age range coverage were observed in Italy (ages 18–64) and Austria (ages 18–45).[10] In addition, with the exception of Bulgaria, the Czech Republic, Estonia, Hungary, and Sweden, all countries excluded the institutionalised population.

[10] Clearly, the deviating age range in Austria and Italy forecloses a comparison of results for these two countries with those of other countries across the whole age range. It only makes sense to include results from Austria and Italy in a cross-national comparison if the age range is restricted to one that matches their range.

Table 2: Fieldwork characteristics of wave 1 of the Generations and Gender Survey, by participating country

		Population coverage		Data collection		
Country	**Pilot?**	**Age range**	**Institutionalized people included?**	**Start**	**End**	**Mode**
Australia	yes	15+	no	08/2005	03/2006	PAPI or Phone, SAPQ
Austria	no*	18–45	no	09/2008	02/2009	CAPI
Belgium	yes	18–79	no	02/2008	05/2010	CAPI
Bulgaria	yes	18–79	yes	11/2004	01/2005	PAPI
Czech Republic	yes	18–79	yes	02/2005	09/2005	PAPI
Estonia	yes	21–80	yes	09/2004	12/2005	PAPI, SAPQ
France	yes	18–79	no	09/2005	12/2005	CAPI
Georgia	yes	18–79	no	03/2006	05/2006	PAPI
Germany	yes	18–79	no	02/2005	05/2005	CAPI
Hungary	yes	21–79	yes	11/2004	01/2005	PAPI
Italy	no	18–64	no	11/2003	01/2004	PAPI
Lithuania	yes	18–79	no	04/2006	12/2006	PAPI
Netherlands	yes	18–79	no	10/2002	01/2004	CAPI, SAPQ
Norway	yes	18–79	no	01/2007	09/2008	CATI, SAPQ, Register
Poland	yes	18–79	no	10/2010	02/2011	PAPI
Romania	yes	18–79	no	11/2005	12/2005	PAPI
Russian Federation	yes	18–79	no	06/2004	08/2004	PAPI
Sweden	yes	18–79	yes	04/2012	04/2013	CATI, SAPQ, Register
Germany-Turks	yes	18–79	no	05/2006	11/2006	CAPI

*The questionnaire was already tested in Germany prior to the Austrian GGS. Instead of a pilot, 30 test interviews were conducted in Austria.

CAPI = Computer-Assisted Personal Interviewing, CATI = Computer-Assisted Telephone Interviewing, PAPI = Paper-and-Pencil Personal Interviewing, SAPQ = Self-Administered Paper Questionnaire

Most countries, except for Austria and Italy, conducted a pilot survey to test fieldwork procedures and the questionnaire. Austria skipped the pilot, as they used the same questionnaire that had already been tested and used in Germany. The surveys for which data are currently available were conducted between 2002 and 2013. In most countries fieldwork took a few months to conclude, but in Belgium, Estonia, the Netherlands, and Norway it took more than a year. Most countries used just one data collection mode. In Eastern and Southern Europe (Bulgaria, the Czech Republic, Georgia, Hungary, Italy, Lithuania, Poland, Romania, and the Russian Federation) an interview with a paper questionnaire was the preferred mode. In Western European countries (Austria, Belgium, France, and Germany) computer-assisted personal interviewing was the preferred mode. Five countries used a mix of methods. In Australia two methods were used as alternatives, with respondents being interviewed by telephone if they could not be interviewed in person; each respondent had to fill in a self-completion questionnaire. In the other countries (Estonia, the Netherlands, Norway, and Sweden) different types of information were collected by different methods. For example, in the Netherlands factual information on life histories and social networks

was collected in a face-to-face interview, but attitudinal information was collected in a supplementary questionnaire that respondents had to fill in by themselves.

Table 3: Additional fieldwork characteristics of wave 1 of the Generations and Gender Survey, by participating country

Country	Min. # contact attempts	Average interview length	Incentives?	Type
Australia	n.a.	1 h 13 min	yes	$25/respondent & $25/household
Austria	n.a.	1 h 4 min	yes	Supermarket cheque (€15)
Belgium	3	1 h 13 min	no	
Bulgaria	3	?	no	
Czech Republic	4	1 h 17 min	no	
Estonia	3−5	1 h 39 min	no	
France	7	1 h 5 min	no	
Georgia	3	1 h 11 min	no	
Germany	4	57 min	yes	Lottery ticket (€10)
Hungary	3	1 h 30−40 min	yes	Small gifts (€1/person)
Italy	n.a.	1 h 10 min per HH	no	
Lithuania	3	1 h 20 min	no	
Netherlands	3−10	1 h 14 min	yes	Gift voucher (€10)
Norway	n.a.	43 min	yes	Gift voucher (€1250) for 7 respondents
Poland	n.a.	?	no	
Romania	n.a.	1 h 30−40 min	no	
Russian Federation	3	1 h 51 min	yes	RUB100−500
Sweden	n.a.	26 min	no	
Germany-Turks	4	1 h 13 min	yes	Lottery ticket (€10)

n.a. = not applicable; ? = no information available

From Table 3 it is clear that at least three attempts were made to contact respondents. The minimum number of contact attempts sometimes varied within countries by contact method. For instance, if a telephone number was available in the Netherlands, at least ten contact attempts at different times of the day had to be made. If no telephone number was available, at least three visits to the address had to be made.

The average length of the interview varied from 26 minutes in Sweden to 1 hour and 51 minutes in the Russian Federation. Across countries, the average length of an interview lasted around an hour and a quarter (72 minutes). However, there were substantive differences across countries. There could be a number of reasons for this variation. Survey mode appears to be an important factor: average interview length was shorter in countries that conducted a computer-assisted interview (57 minutes) than in countries that conducted a paper-and-pencil interview (86 minutes). In addition, some countries included optional sub-modules in the survey or added country-specific questions to the questionnaire schedule.

Finally, Table 3 includes information on the use of incentives to stimulate participation. No specific recommendation on the use of incentives was provided in the fieldwork guidelines. Most countries gave a small incentive to participants, either in

cash or as a voucher. Specific information on the effect of these incentives in the GGS is lacking, but several studies have shown that the use of this type of incentive can enhance participation (Singer et al. 1999; Singer and Ye 2013).

Several conclusions can be drawn from the available information on the GGS fieldwork. Firstly, to a large extent countries stuck to guidelines regarding coverage of the population. Researchers should nonetheless be aware of differences between countries, such as inclusion or exclusion of the institutional population or the use of a different age range. Secondly, researchers should be aware that the timing and duration of the wave 1 fieldwork differed substantially across countries. Observed differences between countries could therefore not only reflect genuine country differences but also partially reflect period differences that operate across countries. Thirdly, specific procedures like choice of survey mode or use of incentives differed across countries. These choices were left up to the country teams and thus reflect country differences in best practices as well as in terms of budgetary opportunities and constraints. There is a growing body of literature (e.g., Klausch, Hox and Schouten 2013; Schouten et al. 2013) showing that the choice of survey mode influences the answering patterns of respondents, and if countries use different modes then observed country differences in answers could reflect mode differences rather than genuine country differences.

4. Response

4.1 Introduction

High survey saturation and declining response rates to survey requests in contemporary European societies represent a serious problem for social science research at large. A report on the rate of nonresponse, cooperation, contact, and refusal must therefore be a constitutive part of the methodological report of any survey. To enhance comparability, a detailed description of the calculation methods and an agreed-upon standard should be used (Lynn et al. 2001). Within the GGS, the AAPOR Standard definitions document (American Association for Public Opinion Research 2011) will be used to calculate response rates (Kveder 2005). This is the leading industry standard for reporting the outcomes of a survey process and the calculation of different response and cooperation rates for telephone, in-person, mail, and web surveys.

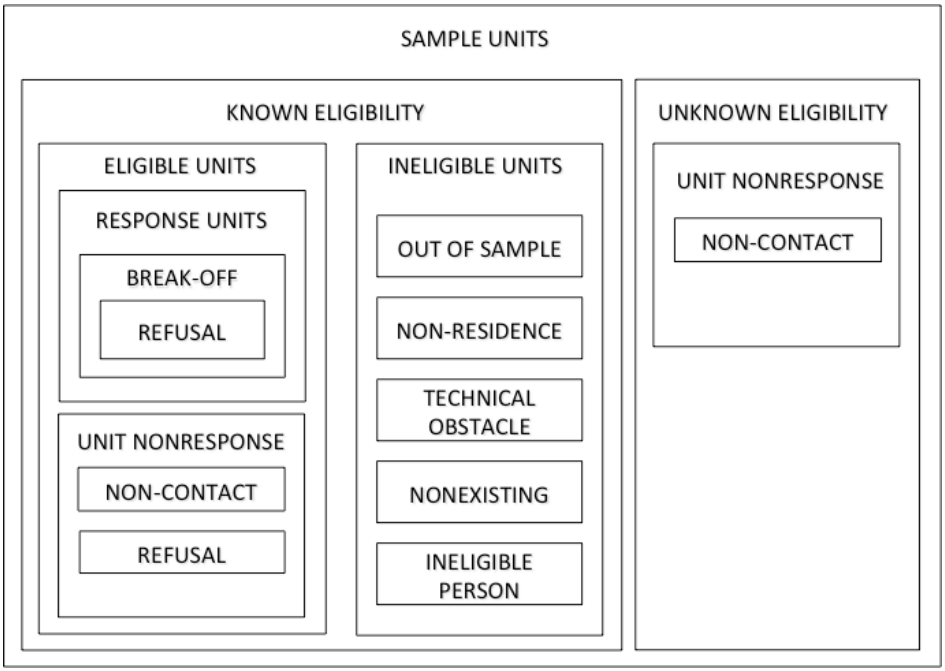
4.2 Types of nonresponse

Survey nonparticipation or unit nonresponse is unwillingness or inability of the potential respondent to share his or her experiences or attitudes in response to a survey request. However, nonparticipation in the survey is not necessarily the result of refusal on the part of the respondent. Inability to establish contact with a sample unit is also considered to be a cause of unit nonresponse.

To calculate nonresponse rates it is necessary to start from a comprehensive definition and classification of all possible outcomes that a sample unit might result in. Figure 1 presents all possible negative outcomes of the sampling unit. The first stage is to establish the eligibility of a sample unit according to the criteria defining the target population of a survey. Most commonly, eligibility refers to residence status and age (e.g., non-institutionalised residents of the Netherlands aged 18–79). Depending on the sample design and sampling frame, eligibility can be established either at the sampling stage or when the initial contact with the sampling unit is successful as part of the fieldwork procedure. The eligibility of the sampled units can be determined at the sampling stage when the sampling is based on a name-list sampling frame such as a central population register. When the sampling design is based on a list of geographic (e.g., settlements, streets, house numbers) or geopolitical units (e.g., census districts), the eligibility of the target unit is usually established after the successful initial contact. Any sample units that could not be contacted are classified as unit nonresponse due to non-contact with unknown eligibility. However, determination of ineligibility does not always require a successful contact attempt. If an address does not exist or is out of primary sampling unit designation, is an empty dwelling or is clearly not a residential building, the sample unit is designated as ineligible.

Once the eligibility of the target individual has been established, an interview is attempted. Refusal to participate can be given either at the start of the contact attempt or later, after the interview has started. Depending on the amount of information collected up to the break-off, a refusal can be classified either as a complete refusal or as a partial interview. The criterion determining whether the break-off is considered a partial interview is part of the survey parameters and is based on the decision of the survey-taking organisation.

Figure 1: Final negative outcomes of a survey request



4.3 Final disposition codes and calculation of response rates

The calculation of response rates starts with assigning a final disposition code to all sampled units in the survey process. Definitions of these codes follow the AAPOR recommendations (American Association for Public Opinion Research 2011). A successful outcome of a contact attempt results in an interview. The outcome is defined as a complete interview (*I*) if all the questions have been answered. If the interview has been started but has not reached completion, a partial interview (*P*) or a break-off (*R*) might be recorded. The criterion as to when to treat the partially completed interview as a break-off is rather arbitrary and based on the decision of the survey organisation. The most commonly accepted rule of thumb is that anything below 50% of completed responses signifies a break-off. Refusal (*R*) represents the decision of the potential respondent not to comply with the survey request. A refusal implies that the contact has been successfully established with an eligible target respondent, but that this respondent is unwilling to participate in the survey.

Not all sample units can be successfully contacted during the fieldwork process. Some of the failed contacts are known to be eligible – their eligibility being determined either from the sampling frame information or during the initial contact with a household member. This group is thus made up of resolved non-contact cases (*NC*). Often, however, eligibility cannot be established beforehand: hence the non-contacted sample unit remains within the group of units with unknown eligibility (*UH*). As it is reasonable to assume that not all the unresolved units are eligible, an expected proportion of eligibility (*e*) needs to be estimated in order to properly account for uncertainty.

The most commonly used calculation of the response rate is the ratio of successful interview outcomes (complete and partial interviews) to all eligible units in the sample. However, the exact number of eligible cases is often unknown and an assumption needs to be made regarding the eligibility of sample units with undetermined eligibility. To this end, the ratio of potentially eligible units among those with unknown eligibility (*e*) needs to be estimated. The ratio *e* can be derived from secondary sources and its estimation needs to be documented. This leads to the following definition of the response rate (designated as RR_4 in the AAPOR Standard Definitions):

$$RR_4 = \frac{I + P}{(I + P) + (R + NC + O) + e(UH + UO)}$$

where I = complete interview; P = partial interview; R = refusal and break-off; NC = non-contact, eligible; O = other, eligible; UH = unknown if household/occupied housing unit – non-contact, unknown eligibility; UO = unknown eligibility, other; and e = estimated proportion of cases of unknown eligibility that are eligible.

For calculation of the response rates in the GGP, the *e* ratio among the undetermined units has been assumed to be the same as among the group with determined eligibility. This can be viewed as a rather stringent assumption, given the likelihood of the eligibility ratio among undetermined units often being lower than among units with determined eligibility. Hence RR_4 can be viewed as a lowerbound estimate of the response rate. We consequently also estimate an upperbound response rate (known as RR_6), where it is assumed that *e* either equals zero (no eligible units among the undetermined) or there are no undetermined cases.

$$RR_6 = \frac{I + P}{(I + P) + R + NC + O}$$

Finally, we also calculate an average rate based on RR_4 and RR_6, which might be more realistic than either of these two:

$$RR_{4-6} = \frac{RR_4 + RR_6}{2}$$

In addition to the overall response rate, a few other rates are informative. The cooperation rate (known as $COOP_1$) is the proportion of all sample units that completed the interview among all those that were contacted during the fieldwork process:

$$COOP_1 = \frac{I}{(I+P)+R+O}$$

The refusal rate (*REF*) represents the clearest negative outcome of the survey process. It is the proportion of sample units that have openly stated unwillingness to cooperate with the survey request. Two rates – REF_2 and REF_3 in the AAPOR Standard Definitions – are used in this calculation and their average is presented:

$$REF_2 = \frac{R}{(I+P)+(R+NC+O)+e(UH+UO)}$$

$$REF_3 = \frac{R}{(I+P)+R+NC+O}$$

The distinction between the two rates is analogous to the distinction discussed when comparing the two calculations of the response rates (RR_4 and RR_6). Finally, the contact rate (*CON*) reflects the ability to locate and contact all designated sample units, and is the proportion of cases where the contact was successful among all the eligible sample units. Again, two rates – CON_2 and CON_3 in the AAPOR Standard Definitions – were used in the calculation and their average is presented:

$$CON_2 = \frac{(I+P)+R+O}{(I+P)+(R+NC+O)+e(UH+UO)}$$

$$CON_3 = \frac{(I+P)+R+O}{(I+P)+R+NC+O}$$

The differences in the specification of the denominator are analogous to the calculations of the response and refusal rates. The contact rate points both to the quality of the sampling frame and to the efficiency of the fieldwork organisation in terms of ability to resolve a set of designated sample cases.

4.4 Response rates in GGS wave 1

Table 4 shows comparative rates estimated on the final disposition codes of 14 GGP countries.[11] There is considerable cross-national variation in the reported response rates. Most of the countries were able to achieve response rates (column *RR*) higher than 50%, with four countries (Bulgaria, Estonia, Georgia, and Romania) even surpassing the 70% threshold. A number of countries had relatively poor response rates (Belgium, Lithuania, the Netherlands, and the Russian Federation), which should be cause for concern. The main reasons for these low response rates are inability to contact the sample units and their unwillingness to cooperate. One exception to this observation is the Netherlands, where the very high refusal rate clearly suggests a hostile survey climate on the one hand and high efficiency in terms of high contact rates on the other.[12]

Table 4: Response, cooperation, refusal, and contact rates of wave 1 of the Generations and Gender Survey, by participating country (in %)

Country	RR_4	RR_6	$RR_{4\text{-}6}$	$COOP_1$	REF	CON
Austria	61.3	67.8	64.6	67.8	25.1	95.2
Belgium	41.8	41.8	41.8	48.3	33.1	86.5
Bulgaria	74.8	81.5	78.1	85.3	13.1	91.6
Czech Republic	49.1	49.1	49.1	53.6	42.4	91.5
Estonia	70.2	70.2	70.2	78.2	15.9	89.8
France	65.2	67.3	66.8	77.0	14.9	86.7
Georgia	71.5	85.0	78.2	89.6	2.8	87.4
Germany	55.4	55.4	55.4	62.0	28.4	89.2
Lithuania	35.6	35.6	35.6	48.8	27.3	72.9
Netherlands	44.6	44.6	44.6	49.1	44.9	90.9
Norway	60.2	60.2	60.2	66.0	26.7	91.3
Romania	83.9	83.9	83.9	97.0	2.6	86.6
Russian Federation	44.8	54.6	49.7	54.6	32.9	91.0
Sweden	54.7	54.7	54.7	66.8	22.3	81.8
Germany-Turks	34.5	34.5	34.5	47.1	24.7	73.3

Note: See text for a definition of these rates

[11] It was not possible to estimate these rates for Australia and Italy, as their data were implemented as an add-on to an existing household panel survey (HILDA and FSS, respectively), nor for Hungary, as the Hungarian GGS wave 1 data are derived from combining data from two national panel waves. In addition, at the time of writing, information about Poland was not available.

[12] It should be taken into account that the Dutch GGP was conceived as a multi-person survey, and some respondents may have refused cooperation in advance because they did not want to involve multiple family members.

The unweighted average response rate across the 14 countries included in Table 4 is just below 60%, so four out of ten respondents were not interviewed. To put this response rate into perspective, it can be compared to averaged country response rates in other cross-national comparative surveys. The Survey of Health, Aging and Retirement in Europe (SHARE) reports an overall wave 1 household level response rate for ten countries of 55% (De Luca and Peracchi 2005, p. 96). In wave 3 of the European Social Survey (ESS), conducted in 2006, the averaged national response rate for 25 participating countries was 64%.[13] Finally, the 2009 wave of the European Union Survey of Income and Living Conditions (EU-SILC) reports a response rate for new households of 73% for 29 participating countries (EUROSTAT 2011, p. 27). Although it is hard to draw definitive conclusions, given differences in sampling frames, survey mode, and survey agencies (e.g., official statistical offices in the case of EU-SILC versus private fieldwork agencies in the case of the academic surveys SHARE and ESS, with the former usually generating a somewhat higher response rate (Groves and Couper 1998)) these figures suggest that the average response rate of the GGS is similar to those of other major comparative surveys in Europe.

5. Weighting

The point of analysing data from a sample is to generalise findings to the target population. To be able to do this, the characteristics of the sample should closely reflect the characteristics of the target population. Factors such as unequal probabilities of selection and differential nonresponse and coverage rates might, however, cause the sample to give a biased representation of the target population. When this is the case a correction is needed, which is usually done by weighting the survey data: individuals that are underrepresented in the achieved sample receive a higher weight, while those who are overrepresented receive a lower weight.

In the GGP Sample Design guidelines (Simard and Franklin 2005) no uniform weighting procedure is recommended. Therefore most of the countries designed their own methods and provided country-specific weights.[14] The only exceptions are Bulgaria, the Czech Republic, Poland, Romania, and Italy.[15]

[13] Retrieved 8/1/2013 from http://ess.nsd.uib.no/ess/round3/deviations.html

[14] Generally, this is variable 'aweight' in a country data file, but variable 'aweight_2402' for Australia and variable 'aweight_1802' for the Netherlands.

[15] The Italian GGP team did provide weights for the original survey data, but these were designed for the original household sample survey. This sample was converted into a person sample, and new weights had to be developed subsequently.

No detailed information about the construction of the country-specific weights is available for most countries.[16] In particular, information is missing on whether or not the country-specific weights are design weights, post-stratification weights, or a combination of the two. It seems most probable to assume that in most countries some sort of post-stratification weighting has been applied. It is therefore important to validate the country-specific weights by examining to what extent the weighted estimates accurately reflect the estimates of the target population (see section 6 – Representativeness). Available information about the characteristics on which weighting has occurred and on the variation in weights is presented in Table 5.

Table 5: Weighting characteristics of wave 1 of the Generations and Gender Survey, by participating country

		Country-specific weight		Centrally constructed weight	
	Weight factors	**Min**	**Max**	**Min**	**Max**
Australia	age, sex, region, employment status, marital status, household composition	0.11	15.17		
Austria	age, sex, labour market participation, country of origin, household type, parity of women	0.29	3.82		
Belgium	age, sex, region	0.79	1.17		
Bulgaria	no country-specific weights			0.66	1.70
Czech Republic	no country-specific weights			0.30	1.55
Estonia	age, sex	0.80	1.49		
France	age, sex, urbanisation, citizenship, social and occupational status, household type, number of household members	0.20	8.69		
Georgia	age, sex, region.	0.58	1.39		
Germany	age, sex, region, education, household type	0.10	8.62		
Hungary	age, sex, region	0.39	3.07		
Italy	no country-specific weights			0.42	2.72
Lithuania	age, sex, urbanisation	0.43	1.84		
Netherlands	age, sex, region, urbanisation, household type	0.19	8.28		
Norway	age, sex, region, centrality, education	0.12	2.11		
Poland	no country-specific weights			0.23	2.55
Romania	no country-specific weights			0.39	1.80
Russian Federation	age, sex	0.46	2.75		
Sweden	age, sex, region, country of birth, education, income, family status	0.41	2.15		
Germany-Turks	age, sex, region, education, household type	0.22	9.38		

[16] To obtain information about practical matters and procedural and technical issues of the GGP in-country implementation and to gain more insight into and improve the quality of the GGS data, the GGP country teams were asked to fill in a metadata grid, an EXCEL file with pre-structured questions. The obtained information has been made publicly available through the NESSTAR system on the GGP web pages. Most of the requested information, however, concerns descriptions of the sampling methods, mode of data collection, fieldwork procedures, and nonresponse; little information was collected on country-specific pre- and post-weighting procedures.

Table 5 shows that the majority of country-specific weights are at least partly determined by age, sex, and either region or urbanisation. This is unsurprising, given the recommendation in the Sample Design guidelines to stratify the sample on these characteristics (see Section 2 – Sample). In addition, age, sex, and region are among the few characteristics for which reliable population data are available in most countries. In some countries weighting was done on additional characteristics, like education (Germany, Norway, and Sweden) and household type (Australia, Austria, France, Germany, and the Netherlands).

Table 5 also reports the range (minimum and maximum values) of the country-specific weights. Extremely low and high weights imply that some population groups are respectively highly overrepresented or practically not represented in the sample. Such weights inflate standard errors, reducing the precision of the survey estimates and causing the weighted sample to be less efficient. In addition, the greater the range of the weights, the greater the increase in variances of the estimates. Biemer and Christ (2008, p. 335) suggest that adjustment factors exceeding 6 are often considered to be too extreme. As Table 5 shows, the range of the weights varies substantially across the country samples. Rather low and/or high maximum values of country-specific weights are found in Australia (min. 0.11, max. 15.17), France (max. 8.69), Germany (min. 0.10, max. 8.62), the Netherlands (max. 8.28), and Norway (min. 0.12). In the other countries the minimum value is 0.19 or higher and the maximum value does not exceed 4.

For those countries that did not calculate personal weights themselves – Bulgaria, the Czech Republic, Poland, Romania, and Italy – we calculated post-stratification weights. For each country separately, these 'centrally constructed' weights adjust the sample distribution for age, sex, region, and household size to the distribution of the target population.[17] More specifically, post-stratification was based on fitting weights to population data on (a) cross-classification of age (four categories for Bulgaria, the Czech Republic, Poland, and Romania: 18–34; 35–49; 50–64; 65–79; three categories for Italy: 18–34; 35–49; 50–64), sex, and region (two NUTS-1 categories in Bulgaria, eight NUTS-2 categories in the Czech Republic, 16 NUTS-2 categories in Poland, four NUTS-1 categories in Romania, and five NUTS-1 categories in Italy) and (b) cross-classification of age (see above), sex, and household size (one-person versus multi-person) using the iterative raking procedure in STATA. Cases with missing values on one or more weight factors have missing values on the post-stratification weight

[17] There are a number of reasons why the construction of the post-stratification weights is restricted to the four socio-demographic variables of age, sex, region, and household size. Firstly, they are important dimensions of research on demographic behaviour. Secondly, these variables are among the few for which population data are generally available for all or most countries, unlike, e.g., education and labour market participation. Finally, we restricted ourselves to the four socio-demographic variables to keep the number of strata small enough to yield satisfactory estimators.

variable too. Moreover, like the country-specific weights, the centrally constructed weights are re-scaled so that the weighted sample size equals the unweighted sample size. As presented in Table 5, the range of the weights for Bulgaria, the Czech Republic, Poland, Romania, and Italy lies at 0.66–1.70, 0.30–1.55, 0.23–2.55, 0.39–1.80, and 0.42–2.72, respectively, thus neatly staying within the boundaries suggested by Biemer and Christ (2008).

6. Representativeness

As discussed in the previous section, to make inferences about the characteristics of populations based on sample data, the sample should be unbiased. In other words, the sample should be representative of the target population, at least in those characteristics most important to the study. To assess representativeness of the GGP wave 1 data, for each country we compared the survey sample to the target population (non-institutionalised population aged 18–79 [18]) for the following socio-demographic characteristics: age, sex, region, marital status, household size and composition, educational level, and unemployment. The rationale behind the selection of these characteristics is their strong connection to different types of demographic behaviour and events, and the public availability of population estimates for these characteristics. It would be preferable to include additional characteristics of a non-demographic nature to assess bias in the national samples on other dimensions as well, but for these types of characteristic no cross-national population-based information is available. We therefore limit ourselves to socio-demographic characteristics.

Analogously to the post-stratification weighting, age is generally split into four age categories: 18–34, 35–49, 50–64, and 65–79.[19] The number of region categories varies across countries,[20] household size differentiates between one-person and multi-person households, and marital status includes the categories of never married, married, divorced, and widowed.[21] National target population data by age, sex, region, and

[18] For the exceptions – i.e., a (slightly) different target population – see Table 2 in section 2 – Sample.

[19] Deviant age categories are: '15–34' and '65 and over' for Australia; '18–34', '35–45' and lack of the two oldest age groups for Austria; '21–34' and '65–80' for Estonia; '21–34' for Hungary; and '18–34' and lack of the oldest age group for Italy.

[20] If possible, region categories correspond to NUTS-1 classifications; otherwise, the categories of variable 'aregion' are used. The latter holds for Australia, the Czech Republic, Estonia, Georgia, Lithuania, Norway, and the Russian Federation.

[21] We also collected data on household composition, distinguishing between five types of private household: persons living alone, lone parents living with at least one child, partnered couples (married or cohabiting) with or without children, children living with at least one parent, and the remainder (other composition in private household). These data were not available for all countries, and will not be discussed in this section. Information on the sample and population distribution of this variable is available at the GGP website.

marital status was obtained from Eurostat and is collected for the survey year (or, if data collection was spread over two or more years, the year in which most of the interviews took place). For most countries, population information of household size was obtained from the 2001 Census.[22]

In addition to this population-based information, data on educational level is taken from the Labour Force Survey for the particular survey year.[23] Although the accuracy of population estimates obtained from the Labour Force Survey or other survey samples is lower than for estimates based on population registers or census data, we decided to include educational level, as it is commonly used as main correlate of demographic behaviour. Education is divided into three levels: up to lower secondary, upper secondary, and tertiary.

The amount of information per country on the representativeness of the data is vast. For each characteristic we have information on the distribution of the population, the unweighted sample, and the weighted sample. We refrain from presenting and discussing all that material in detail.[24] Rather, we focus on an overall indicator of the level of bias in a characteristic in the unweighted and weighted GGS samples. Let P_j be the proportion of the population in category j of a specific characteristic, and p_j be the proportion of the sample in category j of that characteristic. Then we define our indicator of bias as the percentage of weighted deviations between the sample and population categories. This can be expressed as:

$$\text{Bias} = 100 * \sum_{j=1}^{n} (P_j * \left| 1 - (\frac{p_j}{P_j}) \right|)$$

If the sample proportions of a characteristic completely equal the population proportions, the bias is 0%. A bias of 10% indicates that, on average, each sample proportion is 10% higher or lower than that observed for the population. For example, in Germany 46.1% of the sample was male and 53.9% female, whereas the population consists of 49.5% males and 50.5% females. Hence the percentage of males is underestimated by 6.9 % and the percentage of females is overestimated by 6.7%. The bias in this indicator is 100 * ((.495 * |1-(.461/.495)| + .505 * |1-(.539/.505)|)) = 6.8%. In Table 6, bias estimates are presented for six characteristics: age, gender, region, marital status, household size, and educational level. For each characteristic, the first column shows the bias in the unweighted sample and the second column shows the bias in the weighted sample.

[22] The exceptions are Australia (Census of Population and Housing 2006), Austria (Austrian Microcensus 2008 and 2009), and Georgia (Georgian Population Census 2002).

[23] We also calculated unemployment rates for men and women. As for household size, information on unemployment rates for the sample and the population is available at the GGP website.

[24] Country-specific tables including this information are available at the GGP website.

The results for age in Table 6 suggest that the average deviation per category was more than 11% (11.8), with five countries (Romania, Bulgaria, Italy, Poland, and Romania) exceeding 15%. In six countries (Australia, Italy, the Netherlands, Poland, Romania, and the Russian Federation) there is a clear underrepresentation of the youngest age group (18–34). An overrepresentation of the oldest group (65–79) and/or persons aged 50–64, on the other hand, is found in Poland, Romania, Lithuania, the Russian Federation, and Italy. The opposite distortion is observed in Bulgaria, with an overrepresentation of the youngest and an underrepresentation of the two oldest age groups. After weighting, the bias is reduced by more than half (to 5.7%), with only Australia and Poland exceeding it by 10%. In Australia the underrepresentation of the young remains, whereas in Poland there is an overrepresentation of the 35–49 age group and an underrepresentation of the 50–64 age group. All in all, weighting seems to reduce bias in the age structure of the samples to acceptable levels, which is expected, given that age criteria were used in the weighting procedures in most countries.

As in many surveys, the gender distribution of the GGP wave 1 samples is skewed towards women. This holds especially for Austria, Estonia, the Netherlands, and the Russian Federation. Only in Lithuania are men slightly overrepresented. Overall, the bias is 9%. After weighting, the bias is reduced to just 1.5%, suggesting that weighting is very successful in aligning the gender distribution of the sample to that of the population. Again, this is as expected, given that most countries used gender as a criterion in producing weights. The only country where the bias is still considerable, even after weighting, is the Russian Federation (13.2%).

The overall bias in the geographical distribution of the samples is about 9%. There is particularly clear evidence for under- and overrepresentation of various regions in Belgium, the Czech Republic, Estonia, Italy, Norway, and Poland. After weighting, however, the overall bias is strongly reduced to 2.5%, with the largest bias (10.0%) remaining in the Czech Republic and Estonia. Again, this is as expected, given that most countries used region as a criterion in producing weights.

The bias in marital status is quite considerable (12.9%). It is over 15% in six countries (Australia, Belgium, Georgia, Norway, Romania, and Sweden). The direction of bias is less uniform however. Never-married persons are overrepresented in Georgia and clearly underrepresented in Australia, Belgium, Norway, Romania, and Sweden. Widowed people are overrepresented in Australia, Lithuania, and the Netherlands but underrepresented in Belgium, France, and Germany. Divorcees are clearly underrepresented in Georgia, Germany, and Lithuania, and overrepresented in Australia and the Netherlands. There is an overrepresentation of married persons in the Belgian, Norwegian, and Swedish samples. Weighting alleviates these biases only slightly. After weighting the average bias is still 11.9%, suggesting rather large deviations in marital status distributions even after weighting.

Table 6: Weighted average deviation between sample and population distributions, by country and characteristic (in %)

	Age		Gender		Region		Marital status		Household size		Educational level	
Country	NW	W	NW	W	NW	W	NW	W	NW	W	NW	W
Australia	15.5	13.3	9.4	5.2	7.8	1.0	15.1	15.5	10.0	3.2		
Austria	4.4	3.2	20.6	0.0	5.0	0.8	5.1	3.8	3.2	0.8	5.5	7.7
Belgium	8.1	8.3	3.0	0.4	9.9	1.0	17.3	18.7	4.2	4.0	14.5	14.6
Bulgaria	18.7	3.6	6.0	0.2	0.8	0.2			3.7	0.1	11.8	9.8
Czech Republic	5.1	5.0	1.8	0.4	14.7	9.9	6.3	10.5	27.0	15.6	17.7	16.9
Estonia	7.2	3.3	18.2	0.6	9.9	9.7	10.0	9.7	1.0	2.4	8.9	10.2
France	9.6	4.4	10.2	0.2	8.8	3.7	9.6	12.8	21.8	2.6	13.4	11.4
Georgia	9.3	5.7	5.0	0.6	3.7	0.7	20.9	24.0	20.8	21.6		
Germany	5.1	1.7	7.0	0.0	4.3	0.3	8.8	11.4	12.6	9.0	13.6	10.2
Hungary	6.6	4.9	5.2	1.0			8.9	4.0	1.0	2.2	10.4	8.0
Italy	18.7	1.7	6.4	0.2	11.9	0.0	7.3	4.0	2.2	0.0	21.4	21.3
Lithuania	11.9	10.0	7.2	1.4	2.5	1.9	12.1	15.8	16.4	15.2	2.7	2.2
Netherlands	13.5	2.3	15.8	0.0	8.6	2.3	10.9	1.0	16.6	0.2	19.4	18.0
Norway	9.2	6.1	1.6	0.0	23.6	5.5	18.9	16.2	3.4	3.2	11.1	4.7
Poland	23.5	11.2	13.0	1.0	15.6	0.1			14.4	0.0	2.0	4.2
Romania	27.4	2.6	2.8	0.6	1.8	0.0	15.8	12.5	10.0	0.2	10.4	8.4
Russian Federation	13.8	8.7	16.6	13.2								
Sweden	5.6	5.9	2.6	1.2			26.6	18.3			17.2	14.9
Average	11.8	5.7	8.5	1.5	8.6	2.5	12.9	11.9	10.5	5.0	12.0	10.8

NW = non-weighted, W = weighted

The bias in marital status is quite considerable (12.9%). It is over 15% in six countries (Australia, Belgium, Georgia, Norway, Romania, and Sweden). The direction of bias is less uniform however. Never-married persons are overrepresented in Georgia and clearly underrepresented in Australia, Belgium, Norway, Romania, and Sweden. Widowed people are overrepresented in Australia, Lithuania, and the Netherlands but underrepresented in Belgium, France, and Germany. Divorcees are clearly underrepresented in Georgia, Germany, and Lithuania, and overrepresented in Australia and the Netherlands. There is an overrepresentation of married persons in the Belgian, Norwegian, and Swedish samples. Weighting alleviates these biases only slightly. After weighting the average bias is still 11.9%, suggesting rather large deviations in marital status distributions even after weighting.

The bias in household size is just below 11% (10.5). Biases of over 15% are observed in the Czech Republic, France, Georgia, Lithuania, and the Netherlands. In most countries this bias occurs because one-person households are overrepresented. This is particularly the case in the Czech Republic, France, Lithuania, and the Netherlands. Only in the Belgian, Bulgarian, Estonian, and Georgian GGP samples are people living alone underrepresented. Weighting effectively reduces the overall bias by more than half, to 5.0%. Large biases remain only in the Czech Republic, Georgia, and Lithuania. Additional data on household composition (data not shown) suggest an overrepresentation of cohabiting couples with(out) children in Austria, Belgium, Estonia, Hungary, Italy, Poland, Romania, and Norway, while this group is underrepresented in the Czech Republic, France, Georgia, Germany, Lithuania, and the Netherlands. In addition, several GGP wave 1 samples exhibit an overrepresentation of lone parents with at least one child and an underrepresentation of adult children living with at least one parent.

In terms of educational attainment, the overall bias is again in the order of 10%, with particularly elevated levels of bias in Italy (21%), the Czech Republic (18%), the Netherlands (19%), Belgium (15%), and Germany (14%). In most countries lower-educated people are underrepresented and higher-educated people overrepresented. Weighting reduced the bias in some countries (especially Norway), but increased it in others (Austria, Estonia, and Poland). Overall, weighting hardly limits the bias at all (average bias dropped from 12.0% to 10.8%). An overrepresentation of unemployed people was also observed (data not shown).

Overall, these results show that distributional biases – usually in the order of about 10% – are present in the unweighted samples. The extent of the bias varies across countries, but there are no countries without biases on any of the characteristics surveyed in this section. Weighting does lead to a strong reduction in the bias of the sample distribution for sex and region – and to a lesser extent also for age and household size – for the vast majority of the GGP wave 1 countries. This largely results

from the fact that these indicators were used to weight the data in many countries. At the same time, weights generally have no discernible impact on biases in terms of marital status and educational level. In a small number of cases, country-specific weighting even results in a slight exaggeration of the differences with the target population. It should be kept in mind that part of these observed biases might be due to the less-than-optimal sources of population data we used: the 2001 Census data for marital status and household size, with the likelihood of changes between 2001 and the survey year, and the Labour Force Survey for education and unemployment, which does not allow us to ascertain whether the cause of the biases is on the side of the GGP data or the Labour Force Survey data.

The weights constructed by the country teams were used for most countries. Such weights were missing or inappropriate for some countries: hence weights were constructed centrally (see Section 5). To ascertain the quality of the central weighting procedure we also constructed central weights for the countries that had provided country-specific weights, and compared the distributions on key characteristics when using centrally calculated weights with those when using country-specific weights. The results were quite similar (data not shown), which supports the idea that the centrally constructed weighting method, used to construct weights for Bulgaria, the Czech Republic, Poland, Romania, and Italy, performs as well (or as poorly) as the weights produced by countries themselves.

7. Conclusions

The main goals of this article were to describe the main features of the implementation of the GGS in participating countries, and to describe and evaluate the quality of the data collection of the GGS in terms of their cross-sectional representativeness. Attention was paid to key indicators of survey quality such as sampling procedures, fieldwork implementation, response rates, weighting, and representativeness.

The use of a probability sample was prescribed by the sampling guidelines, and it turned out that all countries complied with that prescription. The exact methods by which this was done were country-specific. In most countries a simple random sample was drawn from national registers or census information, but in countries where this was not feasible other randomised methods were employed.

Guidelines for GGP fieldwork implementation were not very strict, so that most countries adhered to country-specific best practices. Two issues warrant specific attention from GGP users. First, the population covered by the survey (particularly the inclusion or exclusion of institutional population and the age range of the sample) differs between countries. In most countries the variation in age range is relatively

limited, and it is suggested that users drop the small numbers of respondents that were outside the defined GGP universe of the 18–79 age range or were living in institutions. Italy and Austria, however, adopted age ranges that were much more restricted, and comparisons that include these two countries should focus on common age ranges only. Second, the timing and duration of the wave 1 fieldwork differed substantially between countries; therefore country differences could partially reflect period differences that operate across countries. Users should take this into account when interpreting differences between countries.

Response rates in GGP countries varied considerably, from a rather low 42% in Belgium to a very high 84% in Romania. On average the response rate was about 60%, which is highly comparable to the average response rates in other major comparative studies that have been conducted by academics, like ESS and SHARE.

Even more important than the level of nonresponse is the bias in the nonresponse. To evaluate that bias, survey data were compared to population data on a range of characteristics. This comparison was also made for both the unweighted sample and the weighted sample. For unweighted data, considerable bias was detected for such characteristics as age, gender, region, marital status, household size, and educational level. When the data were weighted using either the weights developed by the country teams or the centrally developed weights, biases for age, gender, region, and household size were substantially lower. This does not come as a surprise, since these characteristics were often used to produce weights. However, biases with regard to marital status and educational level remained considerable. Users are therefore advised to apply the weights provided in the GGS datasets whenever feasible. At the same time, caution is recommended when the data are used to describe marital status and educational distributions of the population at large. Users should also keep in mind that there may also be biases with regard to representativeness of other important indicators for which no population data are available, thereby threatening the validity of prevalence estimates (for instance, prevalence of divorce) and, to a lesser extent, the validity of model parameters (for instance, the impact of divorce on income) (Groves 1989).

Within an international collaborative data collection effort like the GGS it is extremely difficult to collect data that are entirely comparable across countries. The GGS could nonetheless profit from the growing literature on enhancing data quality in comparative surveys (Harkness et al. 2010). For instance, the experience of recent surveys like ESS and SHARE suggest that high levels of centralisation and standardisation are essential (Pennell et al. 2010). By comparison, these levels were relatively low in the first wave of the GGS, making it hard to judge whether country differences in observed variable scores reflect true differences in these scores across countries rather than differences in survey implementation. Post-data collection

assessments that compare GGS data to population information on demographic outcomes (see Vergauwen et al. (2015) in this Special Collection) and examine the cross-national equivalence of multiple-item data (see Hox, De Leeuw, and Boevé (in preparation) in this Special Collection) suggest that the GGS does fairly well in these respects. Still, a great deal could be gained by a much more upfront approach to harmonisation, e.g., by centralising methods of questionnaire translation and design and by rigorously applying the same survey modes across countries.

This overview of the data quality aspects of the first wave of the GGS suggests numerous implications for future GGS waves. A first implication is to strengthen efforts to increase cross-national comparability. These efforts could include (a) alignment of the timing of waves in the participating countries, (b) more elaborate fieldwork guidelines, (c) more elaborate and uniform reporting of contact and nonresponse information, and, most importantly, (d) the use of exactly parallel questionnaires and survey modes. Another important implication is to enhance efforts to document the representativeness of the realised samples and to improve uniformity in the weighting procedures. Such efforts could profit from better documentation of national weighting procedures and/or the implementation of central weighting procedures.

References

American Association for Public Opinion Research. (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition: AAPOR.

Biemer, P.B. and Christ, S.L. (2008). Weighting survey data. In: De Leeuw, E.D., Hox, J.J., and Dillman, D.A. (eds.). *International Handbook of Survey Methodology*. New York: Lawrence Erlbaum Associates: 317–341.

De Luca, G. and Peracchi, F. (2005). Survey participation in the first wave of SHARE. In: Börsch-Supan, A. and Jürges, H. (eds.). *The Survey of Health, Aging, and Retirement in Europe – Methodology*. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA): 88–104.

EUROSTAT (2011). 2009 Comparative EU intermediate quality report - version 3 July 2011. Luxembourg: European Commission, EUROSTAT.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. Hoboken, N.J.: Wiley & Sons. doi:10.1002/0471725277.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons. doi:10.1002/9781118490082.

Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L.E., Mohler, P.P., Pennell, B.E., and Smith, T.W. (2010). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, N.J.: Wiley & Sons. doi:10.1002/9780470609927.

Hox, J., De Leeuw, E.D., and Boevé, A. (in preparation). Reliability and Cross-National Comparability of Multiple-Item Constructs.

Klausch, T., Hox, J.J., and Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research* 42(3): 227–263. doi:10.1177/0049124113500480.

Kveder, A. (2005). Definitions and documentation of the final disposition codes. In: United Nations Economic Commission for Europe (ed.), *Generations & Gender Programme. Survey Instruments*. New York and Geneva: United Nations: 115–118.

Kveder, A. (2007). Guidelines for survey fieldwork and panel maintenance. In: United Nations Economic Commission for Europe (ed.), *Generations & Gender Programme. Concepts and Guidelines*. New York and Geneva: United Nations: 45–58.

Lynn, P., Beerten, R., Laiho, J., and Martin, J. (2001). Recommended standard final outcome categories and standard definitions of response rate for social surveys (ISER Working Papers).

Pennell, B.-E., Harkness, J.A., Levenstein, R., and Quaglia, M. (2010). Challenges in cross-national data collection. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L.E., Mohler, P.P., Pennell, B.-E., and Smith, T.W. (eds.). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, N.J.: Wiley & Sons: 269–298. doi:10.1002/9780470609927.ch15.

Schouten, B., Van den Brakel, J., Buelens, B., Van der Laan, J., and Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research* 42: 1555–1570. doi:10.1016/j.ssresearch.2013.07.005.

Simard, M. and Franklin, S. (2005). Sample design guidelines. In: United Nations Economic Commission for Europe (ed.). *Generations & Gender Programme*. *Survey Instruments*. New York and Geneva: United Nations: 3–14.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics* 15(2): 217–230.

Singer, E. and Ye, C. (2013). The use and effects of incentives in surveys. *The Annals of the American Academy of Political and Social Science* 645(1): 112–141. doi:10.1177/0002716212458082.

Vergauwen, J., Wood, J., De Wachter, D., and Neels, K. (2015). Quality of demographic data in GGS Wave 1. *Demographic Research* 32(24): 723–774. doi:10.4054/DemRes.2015.32.24.