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

DEMOGRAPHIC RESEARCH

VOLUME 43, ARTICLE 50, PAGES 1461–1494
PUBLISHED 4 DECEMBER 2020
https://www.demographic-research.org/Volumes/Vol43/50/
DOI: 10.4054/DemRes.2020.43.50

Research Material

Evaluating interviewer manipulation in the new round of the Generations and Gender Survey

Eugenio Paglino

Tom Emery

Contents

Demographic Research: Volume 43, Article 50
Research Material

Evaluating interviewer manipulation in the new round of the Generations and Gender Survey

Eugenio Paglino[1]

Tom Emery[2]

Abstract

BACKGROUND
Past research has criticized the quality of the Generations and Gender Survey retrospective fertility and partnership histories. For example, fatigue and learning effects were deemed responsible for distortions in the Generations and Gender Survey in Germany.

OBJECTIVE
We assess the quality of the Generations and Gender Survey for Belarus (GGS-BL) in 2017 to assess whether the new centralized fieldwork system and monitoring procedures are effective in preventing distortions in life history data.

METHODS
We conduct a range of analyses to find evidence of fatigue and learning effects on the part of both interviewers and respondents. Multilevel models, comparison of crucial indicators with other sources, and descriptive analysis of item-nonresponse are used.

RESULTS
In a preliminary analysis, we find no evidence of severe distortions. An in-depth analysis into interviewer and respondent effects reveals some small signs of possible manipulation. However, when assessing the impact of anomalous interviewers on the indicators more likely to be affected, we find no evidence of harm to data quality.

CONCLUSION
The new data collection procedure adopted by the Generations and Gender Survey seems to be effective in preventing detectable manipulation and fabrication. Furthermore, we dismiss the hypothesis that fatigue and learning effects are a source of bias in the collection of life history data.

[1] University of Pennsylvania, USA. Email: paglino@sas.upenn.edu.
[2] Department of Public Administration and Sociology, Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, the Netherlands. Email: tom@odissei-data.nl.

CONTRIBUTION

This paper delivers three key messages: (1) the Generations and Gender Survey for Belarus is a reliable source for retrospective histories, (2) in-field checks are an effective tool to prevent fabrication, and (3) extensive use of inexperienced interviewers does not seem to harm data quality when adequate monitoring and monitoring is in place.

1. Introduction

The Generations and Gender Programme collects cross-national, longitudinal data on family and relationship dynamics. In 2020 the Generations and Gender Survey (GGS, www.ggp-i.org) will begin a new round of data collection using a new operational model to try and ensure high data quality, greater comparability, and a timelier release of data. The quality of some data in the previous round of the GGS was questioned after the inaccuracy of retrospective fertility and partnerships histories in the GGS data for Germany was discovered (Kreyenfeld et al. 2010). These effects were not as apparent elsewhere in the GGS, and they have been significantly examined in subsequent research (Ruckdeschel, Sauer, and Naderi 2016; Vergauwen et al. 2015). Nevertheless, in this paper we examine data from Belarus where the new GGS model of operation was applied and examine whether the previous issues in GGS data collection have been addressed. We assert that the GGS data is now more robust and manipulation and fabrication are closely monitored and prevented.

We build on the analysis of interviewer effects as explored by Ruckdeschel, Sauer, and Naderi (2016). In addressing the critiques provided by Kreyenfeld et al. (2010), they focus on data fabrication and on the possibility that response patterns might differ by the experience of the interviewer. Having acknowledged the presence of serious distortions in the fertility and partnership histories in the German data for the Generations and Gender Survey (2005), the authors look for signals of fatigue and learning effects by testing four key hypotheses and find evidence to support them. The approach used to assess the German GGS is grounded in rational action theory, according to which both interviewers and respondents try to minimise the cost of the interview by reducing its duration or avoiding unpleasant questions. This is often the starting point of analyses of interviewer effects. This theory has limitations, and a more comprehensive analysis of the role of interviewers' motivations in fabricating data is contained in Koczela et al. (2015).

However, the unusually large amount of information collected about the interviewers during the GGS Belarus 2017 fieldwork allows us to account for several characteristics of the interviewers that go beyond a simple rational action theory

approach. For example, we have information about region of origin, religion, education, and gender values. Moreover, given that the GGS contains extensive information about the respondents and covers a large sample (9,996 individuals), we can also control for many characteristics of the respondents and adjust standard errors to account for interviewer clustering, as suggested by Kane and Macaulay (1993). Given the extensive paradata and interviewer data available, this paper can therefore go further than previous analysis of GGS data and examine the role of interviewer effects in greater detail in the context of a more centralized and controlled fieldwork operations model.

2. Background

2.1 Existing research

The primary aim of this paper is to examine whether the specific interviewer effects on life history information identified in the previous round of the Generations and Gender Survey persist in the new fieldwork model deployed by the Generations and Gender Programme. The literature on interviewer effects has focused mostly on the impact of interviewer's characteristics. Attention has been devoted mainly to race-of-interviewer and gender-of-interviewer effects, occasionally integrated in a "Social Distance" framework (Tu and Liao 2007).

Some common patterns have emerged. It appears that gender-of-interviewer effects are stronger for male respondents paired with female interviewers (Flores-Macias and Lawson 2008; Kane and Macaulay 1993). Also, questions about gender values seem to suffer more frequently from a gender-of-interviewer bias, usually in the form of male respondents adopting more gender-balanced or feminist positions when interviewed by women.

Some authors have looked at data-quality measures, trying to understand if certain interviewer characteristics can lead to more reliable data (Benstead 2013; Tu and Liao 2007), especially when the survey covers sensitive topics (Becker, Feyisetan, and Makinwa-Adebusoye 1995; Catania et al. 1996). In such research, the focus is usually on item-nonresponse, sometimes differentiating between 'Don't Know' and 'Refusal.' Some authors have found that education and race/ethnic distance lead to higher item-nonresponse (Lau 2018). Mixed results are obtained for age and gender, suggesting that the specific type of survey, its framing, the socioeconomic context, and the country were the study took place may be important factors in determining the presence and the direction of interviewer effects.

Various theories have been put forward to explain such interviewer effects. Self-Disclosure Theory predicts that survey respondents will be more likely to reveal sensitive

information about themselves when they perceive the interviewer as sympathetic or non-judgmental (Dykema et al. 2012). Deference Theory predicts that respondents who perceive themselves as subordinates with respect to the interviewer may try to adjust their responses to meet what they imagine are the interviewer's expectations (Benstead 2013). Social Attribution Theory predicts that respondents may adapt their responses to match the attitudes they think the interviewer holds (Blaydes and Gillum 2013). Stereotype Threat Theory holds that survey respondents may sometimes feel that the interview is a testing environment where respondents who feel threatened by stereotyping of the group they belong to may suffer emotional distress and experience increased anxiety, possibly leading to various kinds of distortion (Davis and Silver 2003; Aronson et al. 1999; Gallagher and De Lisi 1994).

Some authors adopt a less theory-driven approach, testing for specific hypotheses. For example, Catania et al. (1996) try to understand if male respondents over-reporting their number of partners may be due to a desire to impress the interviewer (what the authors call the "Macho Effect") or to make both parties at ease through a "Magnification of Similarity". Catania et al. (1996) also try to understand if the absence or weakness of interviewer effects among female respondents is due to the fact that women are generally more open and self-disclosing than men, which they call the "Ceiling Effect".

Few authors have paid direct attention to the effects of interviewers' values. Usually the effects of values have been examined through the demographic characteristics of the interviewer, which are assumed to carry a stereotype informing the respondent about the interviewer's attitude toward various issues. For example, Lau (2018) argues that in African countries, respondents may perceive male and highly educated interviewers as more supportive of democracy. In the same way, Flores-Macias and Lawson (2008) argue that in Mexico, respondents may perceive female interviewers as more supportive of gender equality.

Existing research on improving data quality in the GGS has two limitations. First, it relies on the interviewer's limited demographic information. The extensive data collected on interviewers during the Generations and Gender Survey allows researchers to relax this assumption, as the dataset contains direct information about interviewers' values and therefore allows for screening or interview allocation procedures to mitigate interviewer effects. Second, the existing research focuses on the interviewers' impact on the respondent's attitudes and values. While these are included in the GGS, the primary concern in previous data collection has been the accuracy of demographic data where respondents may have been reluctant to report children out of wedlock, previous relationships, or homosexual relationships, or simply wanted to 'get through' the interview and subsequently reported minimal demographic activity.

2.2 Belarus in the context of the Generations and Gender Survey

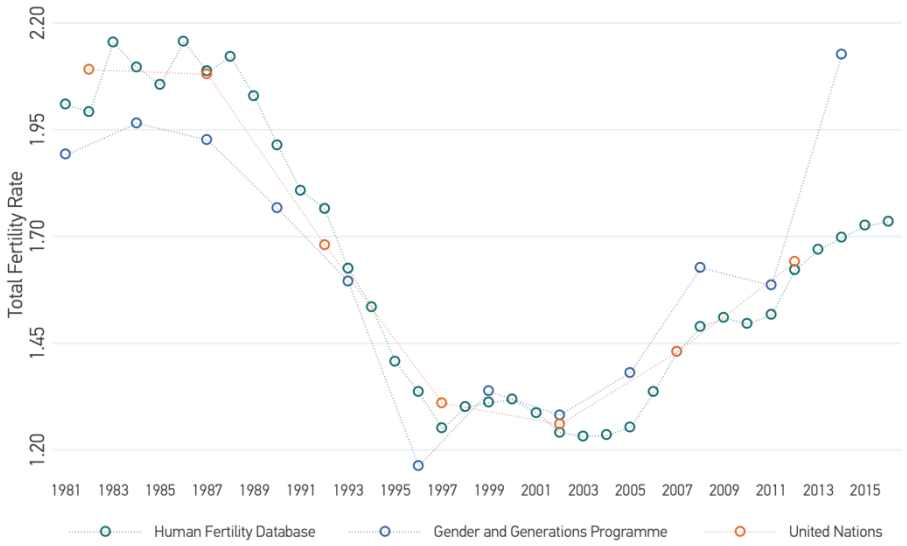
Belarus is used in this analysis as it was the first country to conduct the Generations and Gender Survey using the new centralized fieldwork model, which has been developed to reduce errors, increase monitoring, and improve international comparability of the collected data. In this centralized model of data collection, interviews are collected from respondents and are transferred directly to the central server, administered by the Netherlands Interdisciplinary Demographic Institute (NIDI).[3] This means that the NIDI designs and controls the survey instrument and that data can be checked and verified in real time as the fieldwork is ongoing. This approach is in contrast to fieldwork operations in previous GGS data collection, where each country was responsible for implementing the questionnaire in a bespoke national survey instrument and data was only transferred to the central coordination team at the NIDI upon the completion of all fieldwork (Fadel, Emery, and Gauthier 2020).

However, the use of Belarus as a case study for this analysis is affected by two factors: the demographic context of Belarus and the survey research context of Belarus. Belarus has a total population of 9,507,875 individuals, 77.4% of which live in urban areas (United Nations 2017), and a Gross National Income (GNI) per capita of $5,280 (World Bank and OECD 2017). The Total Fertility Rate (TFR) is 1.73, which is quite high compared to the other countries in the region and to the average for the European Union of 1.57 (Klüsener, Jasilioniene, and Yuodeshko 2019).

The Total Fertility Rate (TFR) in Belarus for the period 1977–2016 (Figure 1) forms a U-shape, with the minimum around 1997, a slight increase afterwards, a decrease again between 2002 and 2006, and a more sustained increase starting from 2011.

[3] NIDI is the Central Coordinator of the Generations and Gender Programme

Figure 1: Total fertility rate in Belarus



According to the analysis in Shchurko (2017), the active role that the Belarussian government started to play in gender education from 2011 (when gender education was legalized) might have influenced this pattern. The gender education policy actively supports and encourages traditional gender roles through the education system. The work of Shchurko (2012) and Polagse (2013) suggests that the government is promoting a conservative view of gender roles, where men should be the breadwinners and become courageous, responsible, and protective fathers, while women should fulfil their 'natural' and 'biological' role as mothers and housekeepers. Attention is also devoted to the promotion of heteronormativity, viewing heterosexual relationships as 'normal' and all other nonconforming gender identities as a threat. Complementary to heteronormativity is pronatalism. A content analysis of government-related media reveals that by condemning both early and late motherhood, specific timing of birth and family size are encouraged, while at the same time suggesting the desirability of having multiple children (Shchurko 2012). The TFR evolution we see in the data is compatible with an effect of governmental gender education on fertility. However, we should beware of interpreting this correlation causally. A deeper analysis is needed, but this is not the concern of this article (Klüsener, Jasilioniene, and Yuodeshko 2019).

The survey research context of Belarus is also of considerable interest regarding the aim of this analysis. Compared to many European Union Member States, Belarus has relatively few social surveys and its market research sector is relatively underdeveloped. The falsification identified in Germany in 2005 was part of fieldwork conducted by TNS Infratest, which is part of a large multinational fieldwork agency, responsible for many large-scale social surveys and with considerable experience in both commercial and scientific surveys. Such agencies do not operate extensively in Belarus. Of the major large-scale international comparative surveys, Belarus only participates in the European Value Study/World Value Survey.

The fieldwork agency used in the European Value Study is the same as that used in the Generations and Gender Survey and is based at the Belarussian State University and headed by Professor David Rotman. The interviewers are generally less experienced and have received less training than interviewers employed by commercial agencies in European Member States. They are paid approximately $8.4 per interview and many are students who conduct the work during their studies. Given these circumstances, it might be expected that the issues identified in the previous round of the Generations and Gender Survey are more likely to present themselves.

Looking at interviewer demographics and activity status, two main groups can be identified: (1) young students below 23 years old, and (2) older and more experienced interviewers. The rational action model does not clearly predict which group will perform better in the context of the Generations and Gender Survey. Few questions in the Generations and Gender Survey could be characterised as sensitive; thus it is not clear what advantages interviewer experience might bring in terms of data quality. However, a well-trained interviewer might learn to recognise strategic misreporting or fatigue on the part of the respondent and might also know how to elicit truthful answers. On the other hand, an experienced interviewer may also know how the external quality checks work and how to shorten the duration of interviews or fabricate data without being discovered. In terms of incentives, a professional interviewer has more to lose if they get caught (they might be fired and not be able to find a new job because of their bad reputation). Nevertheless, the probability of being caught is likely to decrease with experience; thus the expected damage from being caught may be lower for this category. The question then becomes: Are experienced interviewers collecting data of higher quality?

2.3 Hypotheses

Adopting the view that interviewers and respondents are rational agents, we do expect several types of interviewer and respondent effects, especially in terms of falsification or

fabrication of data. To test for the presence of such effects in the data we present the following hypotheses, aimed at giving us specific predictions.

The first hypothesis stems from rational action theory. We expect interviewers to learn how to effectively shorten the interviews as the fieldwork proceeds. By falsely recording that respondents have fewer children, partners, or household members, the interview is shortened considerably without the knowledge of the respondent. We believe that the order of interview is a good proxy for interviewer experience within the Generations and Gender Survey.

H1: Interviewer learning effect: the number of reported children, partners, and other household members decreases, on average, with the order of interview.

The second hypothesis comes from considerations regarding the length and complexity of the Generations and Gender Survey, which might lead to fatigue effects on the part of both the interviewer and the respondent. Fatigue effects might be a consequence of the respondent feeling tired and deciding to underreport specific individuals covered by the survey, such as the number of other household members. This is either because they will have learned that many follow-up questions will be asked for each one, or a consequence of the interviewer perceiving the respondent's fatigue and deciding to misreport the number to reduce the duration of the interview. The series of questions on 'Other Household Members' comes after a similar battery of questions on 'Former Partners' and 'Children'. Given the similarity in question structure, respondents may anticipate that reporting a high number of these individuals will lead to significant follow-up questions and therefore choose to underreport these figures. Underreporting could also be a consequence of the interviewer simply feeling tired and deciding to unilaterally reduce the duration of the survey. A combination of these explanations is also possible.

H2: Interviewer and respondent fatigue effects: the reduction described in H1 will be stronger for other household members, as it is the last of the enumeration questions.

The existing literature has shown that more experienced interviewers are better at collecting survey data (Olson and Peytchev 2007; Lipps and Pollien 2010). This is possible if experienced interviewers grow attached to the institution they work for or if they care about the quality of the data they have collected.

H3: Experience predicts better quality: the order-of-the-interview effect will be stronger for inexperienced interviewers, while in general inexperienced interviewers will

report a lower average number of children, partners, and other household members because of a higher incidence of falsification/fabrication.

From the interviewer's point of view, miscoding nonresponses might be an effective strategy for shortening the interview as it probably requires less planning effort and it is more difficult to detect unless it is systematic. This might be especially true for inexperienced interviewers that do not know the routing of the questionnaire and thus cannot think of more sophisticated forms of fabrication. At the same time, nonresponse may also be a good shortening strategy for respondents. For this reason, we expect a higher incidence of nonresponse in later sections of the GGS, when the respondent is tired and/or annoyed by the interview.

H4: Item-nonresponse will increase with the order of interview and will be higher among young interviewers. It will increase proportionally in later sections of the GGS due to a fatigue effect on the part of both the respondent and the interviewer.

These four hypotheses do not cover all possible interviewer effects. They are instead concentrated on those effects observed in the previous round of the Generations and Gender Survey that are of key concern to the research community looking to use the new round of the Generations and Gender Survey to study retrospective life histories.

3. Data

3.1 Fieldwork

In the present study we use data from the Generations and Gender Survey in Belarus that was conducted in 2017 and data collected from interviewers involved in the fieldwork. Information in this section is drawn from the metadata available on the GGP website (https://www.ggp-i.org/data/browse-the-data). The fieldwork started on April 14th 2017 and ended on the 20th November of the same year. All seven regions of Belarus (Brest, Gomel, Grodno, Mogilev, Minsk, Vitebsk, and Minsk-City) were covered. The target population was the non-institutionalized population aged 18–79 on April 14th 2017.

The sampling frame consisted of a list of household addresses from the 2009 national census, covering 7,359,981 persons aged 18–79 years. A probability sampling method was used. At every selection stage, units were selected based on probabilities proportional to the population size. The first sampling stage consisted of 96 urban and rural survey points, the second stage consisted of 903 census enumeration points, and the third stage of 12,500 households. Within households a single respondent was selected

using the next-birthday method. The target net sample size was 10,000 persons. The final response rate was 76%.

The data was collected in face-to-face interviews using Computer Assisted Personal Interviewing (CAPI). In terms of response rate, there were 119 cases where the address did not exist, 273 cases where no eligible person lived in the household, 1,748 refusals to participate, 55 cases where the selected respondent was unable to answer, and 17 interrupted interviews.

To conduct the fieldwork, a total of 424 interviewers were trained and 333 of these took part in the fieldwork. The organisational structure was as follows: 1 head of the network of interviewers, 7 regional curators, and 14 supervisors. All trained interviewers attended a 5–6 hours' workshop, received appropriate learning material, and were tested to check their skills before they were sent into the field. Each interviewer conducted an average of 30 interviews, with a standard deviation of 37. The mean duration of the interviews was 51 minutes.

Interviewers were payed after their interviews had been correctly uploaded to the server and had undergone quality controls. A payment amounting to 8.4 USD per interview was made by the accounts department of Belarusian State University in the local currency. The interviewers were primarily compiled of two distinct groups. Around 25% were aged over 23 and were known to the Belarussian State University through previous fieldwork data collection projects. Given the size of the GGS however, further interviewers were needed, and the university recruited students to fill this shortfall, who were generally aged under 23 and working on their first fieldwork project. The difference in performance between these two groups is therefore of interest.

Interviewer surveys were conducted toward the end of the fieldwork, from November 10th 2017 to November 19th 2017, and were sent to interviewers who had conducted more than 10 interviews. Responses were collected for 146 out of 236 eligible interviewers, a response rate of 72% representing 71% of all interviews that were conducted. For those interviewers that responded to the interviewer questionnaire we know the age, gender, region and country of origin, whether they are from a rural or urban area, their education level, activity status, marital status, number of children, religion, and responses to section 11 of the Generations and Gender Survey. Section 11 includes two general questions, one about fairness and the other about trust, and five groups of questions regarding (1) general values, (2) which tasks should be performed by society and which by family, (3) care values, (4) intergenerational values, and (5) gender values.

Three levels of quality check were implemented: (1) in-field checks, (2) national team checks, and (3) central team quality checks. The survey itself contains several soft checks to prevent the interviewer from entering erroneous values and corrective instructions were issued when any problematic behavior was identified. For example, if an interviewer tried to input dates of birth which implied that the respondent had a child

when they were under the age of 13, the software would ask for verification before proceeding.

In-field checks were mainly aimed at identifying and correcting mistyping, especially regarding timing of events in relation to a respondent's age. Under the supervision of the national team, an independent team checked 10% of the interviews through telephone call-backs to ensure that interviewers completed the interview in accordance with fieldwork guidelines and recorded all information correctly. This was done by a team at Belarussian State University, who rang respondents to verify that they had indeed completed the interview and that the most important demographic characteristics of the household were correct.

In addition to these two levels, the Central Coordination Team of the GGP performed real-time quality checks on the incoming data, aimed at detecting any anomaly in the data collection process. These included checks for response bias, systematic shortening of interviews, and systematic underreporting of children, other household members, and partners. Warnings were issued whenever an anomaly was detected to ensure that the fieldwork proceeded correctly and that falsification or fabrication of data was prevented. The Central Coordination Team of the GGP provided a weekly report to the Belarussian State University that identified interviewer IDs for which data anomalies were evident. The Belarussian State University team then contacted individual interviewers and/or regional supervisors about the anomalies to discuss rectification of interviewer behavior. This oversight process was most intensive during the first few weeks of the fieldwork, with the aim of demonstrating oversight to both interviewers and regional supervisors.

3.2 Representivity

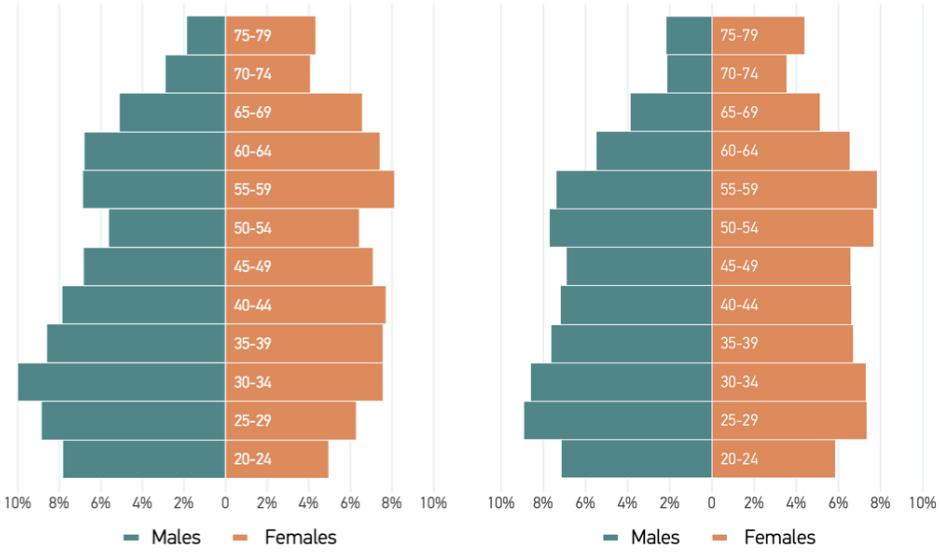
In this section we compare figures extracted from the GGS Belarus 2017 with those from other sources to see if there is any evidence of relevant distortions at the aggregate level, before attempting to identify the interviewer-specific effects that are the primary aim of this paper.

In Figure 1 we present TFR estimates from three different sources: (1) the unweighted GGS Belarus 2017, (2) the Human Fertility Database (MPIDR and Vienna Institute of Demography 2018), and (3) the United Nations World Population Prospects (United Nations 2017). There appears to be no systematic difference across sources, except for the very last period where the GGS overestimates the most recent fertility levels. The high GGS estimate of TFR for this period is a common consequence of the differential response rate of parents and childless individuals in the 18–24 age bracket in a country like Belarus. Individuals without children in this group are likely to temporarily

migrate to neighboring countries either for work or for study and thus have a low probability of entering the sample. Conversely, young adults who already have children are much more likely to be contactable for such a survey. This issue could be fixed with appropriate weights that incorporate parity, but such weights were not applied in this context as it would retrofit the data to the existing TFR trend.

In Figure 2 we compare the population pyramid obtained from the United Nations (2017) with the one from the Belarussian Generations and Gender Survey Sample. Males in the age groups 30–34, 65–69, 70–74, and 75–79 appear to be overrepresented, while those in the 50–54 age group are underrepresented. The distribution seems more even for women, who are only overrepresented in the 65–69 age group and are underrepresented in the 50–54 age group.

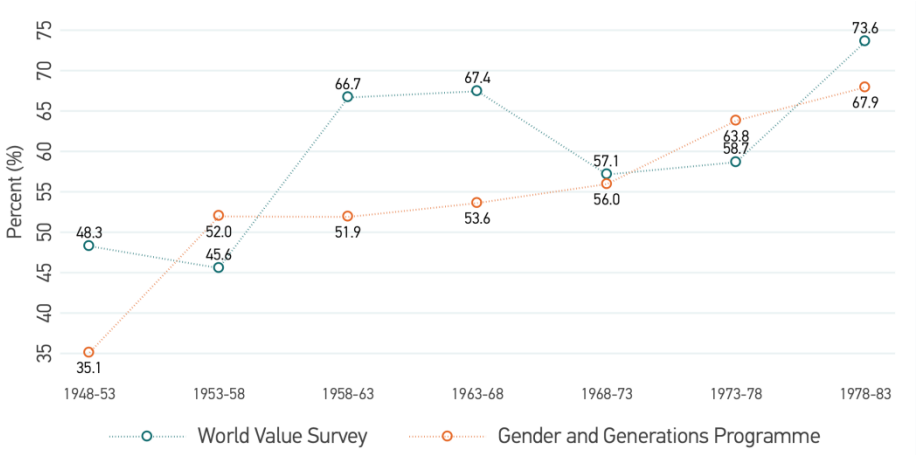
Figure 2: Population pyramids for Belarus based on GGS and UN population data



Finally, we use data from Wave 6 of the World Value Survey (WVS) (Inglehart et al. 2014) as a benchmark for two key statistics for assessing the accuracy of GGS retrospective fertility and partnership histories: the proportion of married women and the proportion of childless women. The reliability of these two statistics was contested in the GGS Germany 2005.

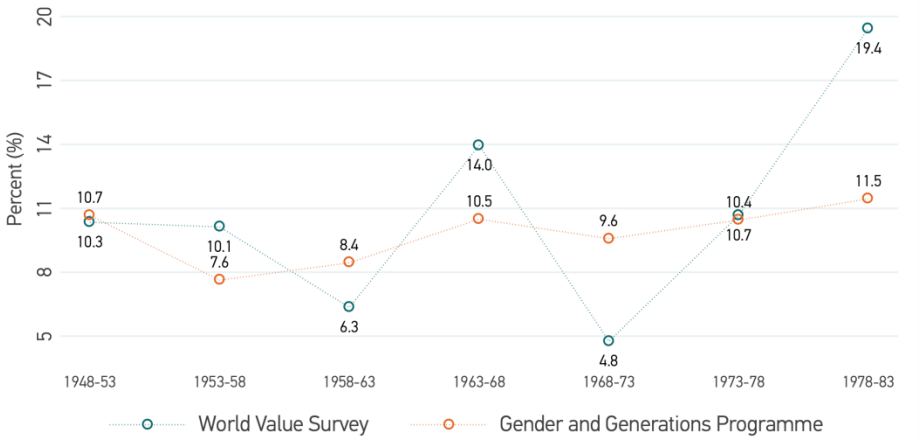
Figure 3 shows the proportion of married women for different cohorts in the GGS and in the WVS. Since Wave 6 of the WVS in Belarus was conducted in 2011, whereas the GGS was conducted in 2017, the composition of the first and last cohorts would have been very different across the two surveys and the comparison would not be reliable, so for this reason we dropped the cohorts 1938–1948 and 1983–1993.

Figure 3: Proportion of married women by cohort in the WVS and GGS



We see no systematic differences between the two sources. We observe a large discrepancy for the two cohorts 1958–1963 and 1963–1968; however, the WVS data seems to vary more between cohorts. Based on existing research and data, the proportion of married women should not change dramatically from one cohort to the next (Tikhonova 2004). Figure 4 shows the proportion of childless women for different cohorts in the two surveys. Compared to the proportion of married women, here the two statistics are closer, and again we see no systematic difference. Also, in this case the data from the WVS has a greater variance due to the small number of cases, which supports the view that the WVS may not be as suitable for cohort analysis as the GGS due to the larger sample size (9,996 for the GGS versus 1,535 for the WVS).

Figure 4: Proportion of childless women by cohort in the WVS and GGS



We similarly compare the proportion of women by parity and marital status. We find very similar figures across WVS and GGS. The results are reported and commented on in Appendix 1.

4. Methods

Our aim is to understand whether there are signs of data fabrication or falsification when looking at the data collected by each specific interviewer.

To test Hypotheses 1 to 3 (hereafter H1, H2, and H3), we start by running a set of linear regressions using as dependent variables: (1) number of 'Non-Applicable' nonresponses; (2) number of 'Refusal' nonresponses; (3) number of 'Don't Know' nonresponses; (4) number of biological children; (5) number of partners, and (6) number of other household members.

We use the order of the interview as the independent variable. It takes value 1 if the interview was the first to be conducted by an interviewer, two if it was the second, and so on. We control for respondent's and interviewer's sex and the interaction between them, respondent's and interviewer's age, a variable for age distance between interviewer and respondent,[4] the total number of interviews conducted by the interviewer, the

[4] We have coded this variable 'Younger' if the respondent is more than 5 years younger than the interviewer, 'About the Same Age' if the age difference is between +5 and –5, and 'Older' if the respondent is more than 5 years older than the interviewer.

respondent's total income,[5] and a binary variable for respondent and interviewer having the same region of origin. As suggested by Kane and Macaulay (1993), we adjust standard errors for interviewer clustering.

We acknowledge that the results coming from such a model cannot be interpreted causally because interviewers are not randomly assigned to respondents; thus true causal effects and compositional effects are mixed and cannot be disentangled. However, we believe that even if an experimental approach were possible, it would not be the best tool to answer our research questions. An experimental approach would force us to focus on an effects-of-causes framework, whereas the study of interviewer effects is more appropriately framed as a causes-of-effects problem. Indeed, from a practical point of view it does not make sense to disentangle the effect of each separate interviewer's characteristics, because in real fieldwork an interviewer is not simply the sum of their characteristics but an individual whose effect on the respondent depends on the rapport between them, the characteristics of both, and the unpredictable outcome of their interaction.

As a robustness check we run a further series of OLS regressions where we add more controls (for example, respondent's region of origin, education, activity status, religion, and marital status). More details about this specification are provided in Appendix 2. In general, the results obtained using these models do not differ substantively from those obtained using the simpler models.

We also try a different approach, used in Ruckdeschel, Sauer, and Naderi (2016), which aims at identifying anomalous interviewers. We run three sets of 333 regressions each (one for each interviewer) with the number of biological children, the number of partners, and the number of other household members as dependent variables. We use the order of the interview, the respondent's age, and the respondent's sex as independent variables. We want to identify those interviewers for which the order of the interview has a significant effect on at least one of the three key variables for shortening the survey, and to investigate their workload in more depth.

$$\boldsymbol{KeyVariable_{r,i} = \alpha_0 + \beta_1 InterviewOrder_{r,i} + \beta_2 Age_r + \beta_3 Sex_r} \qquad (1)$$

The model is specified as in **(1)**; r is one of the respondents and i the interviewer under analysis. $InterviewOrder_{r,i}$ is a set of dummies for the order of interview (first interview, second interview, and so on), Age_r is the respondent's age, and Sex_r is the respondent's sex.

To isolate interviewers with particularly anomalous figures from those who simply happened to interview respondents with, say, a lower average number of children, we

[5] Computed using the information in section 10 of the GGS.

flag interviewers for which the order of interview has a significant effect on at least two of the three screening questions we are considering. We analyze the workload of these interviewers more carefully to assess what its impact might have been on the demographic data produced by the Generations and Gender Programme.

This in-depth analysis also allows us to test H3 and H4 by looking at the group to which the flagged interviewers belong. If H3 and H4 hold, we would expect to find a higher proportion of young inexperienced interviewers among the anomalous ones.

Since we expect to see stronger fatigue and learning effects on the number of other household members, we benchmark the household size distribution obtained from the GGS with the one obtained from the 2009 Census. If our hypotheses are correct and we do find evidence of manipulation at the interviewer level we would expect to have a strong underrepresentation of households with many members in the GGS data.

Finally, to investigate the fatigue effect on the part of the respondent (H4), we look at item-nonresponse by section. If the fatigue effect is strong we expect later questions to have a higher proportion of item-nonresponse. All the analyses are performed using Stata 14.

5. Results

We expect fabrication to occur mainly in the form of shortening the interview, realised by underreporting the number of children, partners, and other household members and thus avoiding follow-up questions. We expect fabrication of this form to be more prevalent in later interviews realised by young and inexperienced interviewers. Moreover, because of a combination of respondents' learning and fatigue, we expect underreporting to be more severe for the number of other household members, because this is the last of the three screening questions asked in the GGS. By looking at some key figures, we try to assess the impact of respondent and interviewer effects on the overall data quality.

5.1 Interview order and fatigue effects in a multilevel model

To test H1 and H2, we start by investigating the effects of the order of interview on the number of 'Non-Applicable',[6] 'Refusal', and 'Don't Know' nonresponses, separately and then aggregated. The results of the multilevel analysis are presented in Table 1. The order

[6] 'Non-Applicable' is offered as an option in some of the questions in case the routing has not worked perfectly and/or the respondent feels they have been asked an inappropriate question that they are unable to answer.

of interview has a significant effect only for the number of 'Don't Know' nonresponses. Moreover, the size of the effect is small and the effect is not significant at the 1% level. Overall, it seems that if an interviewer learning or fatigue effect exists, it does not reveal itself through item-nonresponse.

To see if there is evidence of interviewer learning effects on the three screening questions we run the same model, now using as dependent variables: (1) the number of children, (2) the number of partners, and (3) the number of other household members. The results can be seen in Table 2. The order of interview has a significant negative effect on the number of partners and on the number of other household members. This evidence supports H2, since we find an effect only for the second and third screening questions and not for the first one (number of children).

We obtain similar results with the model using the additional controls mentioned in the methods section. We also perform an extra robustness check by running a model where the effect of interview order can be nonlinear. We find that the second-order term is significant only for the number of refusals, the number of children, and the number of other household members. The third-order term is never significant.

For the number of refusals and the number of other household members, the first-order term is negative whereas the second-order term is positive. For refusals, this could be a sign of a positive learning effect on the part of interviewers. For other household members, it could instead be either a sign of negative learning on the part of interviewers or it could be due to the sampling procedure. Individuals living alone have a lower probability of being sampled when household sampling is used. They may not be at home very much and are thus more difficult to contact. Therefore, it is likely that only a few individuals living alone are part of the sample at the start of the fieldwork. As the fieldwork proceeds, more will be included, thus creating a correlation between the number of other household members and the interview order. This issue is not necessarily problematic, because the two effects (sample selection and bad learning) go in the same direction. Since the overall effect that we found is small, we can be more confident about the absence of substantial negative learning effects.[7]

[7] The same reasoning holds for the number of partners.

Table 1: Descriptive statistics

		Percentage of respondents	Percentage of interviewers
Age group			
	15–19	1.27	30.56
	20–24	6.54	45.83
	25–29	8.78	6.25
	30–34	10.10	7.64
	35–39	9.86	2.08
	40–44	9.64	2.08
	45–49	8.08	2.08
	50–54	7.43	0.00
	55–59	10.59	1.39
	60–64	9.81	0.69
	65–69	8.89	0.69
	70–74	4.11	0.00
	75–79	4.92	0.69
Region of birth			
	Brest	15.33	18.98
	Vitebsk	18.50	15.33
	Gomel	12.85	16.79
	Grodno	11.78	6.57
	Minsk	19.76	13.14
	Mogilev	11.99	13.87
	Minsk-City	9.79	15.33
Sex			
	Male	41.38	22.22
	Female	58.62	77.78
Education			
	Less than University	69.93	33.10
	University or Higher	30.07	66.90
Income (€ per month)			
	0–299	25.76	Not available
	300–399	28.07	
	400–599	24.74	
	600–999	15.97	
	1000–12000	5.46	
Number of children			
	0	21.16	Not available
	1	27.21	
	2	40.60	
	3+	11.03	
Number of partners			
	0	12.58	Not available
	1	75.40	
	2+	12.02	
Number of other household members			
	0	80.04	Not available
	1	10.56	
	2+	9.40	
Number of interviews			
	0–30	Not applicable	66.37
	30–50		11.71
	50–75		15.02
	75–150		4.80
	150–310		2.10
Sample size		3,719	146

Table 2: Multilevel regression of item-nonresponse

	Num. of Non-Applicables			Num. of Refusals			Num. of Don't Knows		
Age of respondent	0.025	*	(0.011)	–0.030	**	(0.010)	–0.005	***	(0.000)
Interviewer's age	–0.003		(0.053)	–0.023		(0.019)	0.002		(0.002)
Total income	–0.000		(0.000)	–0.000		(0.000)	–0.000		(0.000)
Order of the interview	–0.002		(0.003)	–0.002		(0.002)	0.001	*	(0.000)
Number of interviews	0.000		(0.005)	0.001		(0.002)	–0.000		(0.000)
Same region of origin?	0.048		(0.731)	0.775	*	(0.325)	–0.006		(0.025)
Female (R)	–0.999		(0.635)	0.499		(0.356)	–0.029		(0.024)
Female (I)	–0.635		(1.157)	0.429		(0.426)	0.054		(0.033)
Female (R) # Female (I)	0.213		(0.696)	–0.184		(0.396)	–0.029		(0.033)
Young student	0.377		(1.009)	–0.444		(0.519)	0.012		(0.029)
Younger	0.743		(0.635)	0.967		(0.682)	0.058		(0.123)
Older	0.440		(0.374)	0.543		(0.334)	–0.025		(0.042)
Constant	6.264	**	(1.874)	3.160	**	(0.962)	0.279	***	(0.071)
Observations	3,719			3,719			3,719		
Adjusted R-squared	0.012			0.022			0.032		

Note: Standard errors in parentheses. * p<0.05, ** p<0.01. *** p<0.001.

For the number of children, the effect of the first-order term is positive but not significant, while the effect of the second-order term is negative and significant. Positive learning effects on the part of interviewers could explain this, or it could be taken as a sign of the effectiveness of the monitoring system. The first-order term could be telling us that strategic misreporting was reduced over time. The negative sign of the second-order term could be the consequence of a gradual approximation to the true distribution.

Table 3: Multilevel regression of screening variables

	Num. of children (Biol.)			Num. of partners			Number of other household members		
Age of respondent	0.016	***	(0.002)	0.002	*	(0.001)	–0.008	***	(0.001)
Interviewer's age	0.009	**	(0.004)	0.005	*	(0.002)	–0.009	***	(0.003)
Total income	0.000		(0.000)	0.000		(0.000)	–0.000		(0.000)
Order of the interview	–0.000		(0.000)	–0.001	***	(0.000)	–0.002	**	(0.001)
Number of interviews	0.000		(0.000)	0.000	*	(0.000)	0.000		(0.000)
Same region of origin?	0.104		(0.055)	–0.032		(0.033)	0.051		(0.033)
Female (R)	0.103		(0.094)	0.056		(0.054)	–0.365	*	(0.160)
Female (I)	0.059		(0.098)	–0.015		(0.051)	–0.050		(0.159)
Female (R) # Female (I)	0.031		(0.105)	–0.024		(0.057)	0.362	*	(0.169)
Young Student	–0.152		(0.078)	–0.103	**	(0.035)	0.095		(0.072)
Younger	–0.243	*	(0.122)	–0.082		(0.063)	–0.231		(0.178)
Older	0.445	***	(0.079)	0.203	***	(0.051)	0.526	***	(0.126)
Constant	–0.017		(0.151)	0.713	***	(0.082)	3.872	***	(0.207)
Observations	3,719			3,719			3,719		
Adjusted R-squared	0.17			0.044			0.170		

Note: Standard errors in parentheses. * p<0.05, ** p<0.01. *** p<0.001.

5.2 Assessing the impact of interviewer effects

From Table 2 we see that there is an effect of the order of interview on the number of other household members; the effect is also present in the more robust model described in Section 5. However, in both cases the size of the effect is small. To investigate what the impact of this effect on the data quality might be, we can compare the household size distribution obtained from the GGS data with the most recent official statistics available for Belarus in the 2009 Census. If we have a serious underreporting of other household members, we should find a smaller proportion of complex households than in the census. We perform this analysis at the regional level (to get finer information from using the GGS data).

The comparison shows that the GGS underestimates the proportion of individuals living alone and overestimates the proportion of households with two members (Appendix 2 contains the results of this analysis). If we look at households with more than two members the differences become small and the GGS has a slightly higher proportion for all sizes in almost all cases. This last finding might be due to the different criteria used to compute the number of household members in the GGS and the census. The GGS computes the number of household members by counting the respondent, their partner, and their children if living together, and all other persons listed by the respondent when asked about other household members, thus prompting information about specific groups more intensively than the straightforward census question about household size. Overall, we can say that neither the order of interview effect nor the fatigue effect (if present) bias the household size distribution significantly.

5.3 Detecting anomalous interviewers

We also explore an alternative analytical framework to assess the presence of order-of-interview effects in the GGS data. Following the strategy adopted by Ruckdeschel, Sauer, and Naderi (2016), we run three separate regressions for each interviewer with the following dependent variables: (1) the number of children, (2) the number of partners, and (3) the number of other household members. Due to the limited sample size, which in each regression is equal to the total workload of the interviewer, we cannot include many controls. We opt for age and sex of the respondent as the most notable and pervasive covariates across the three dependent variables.

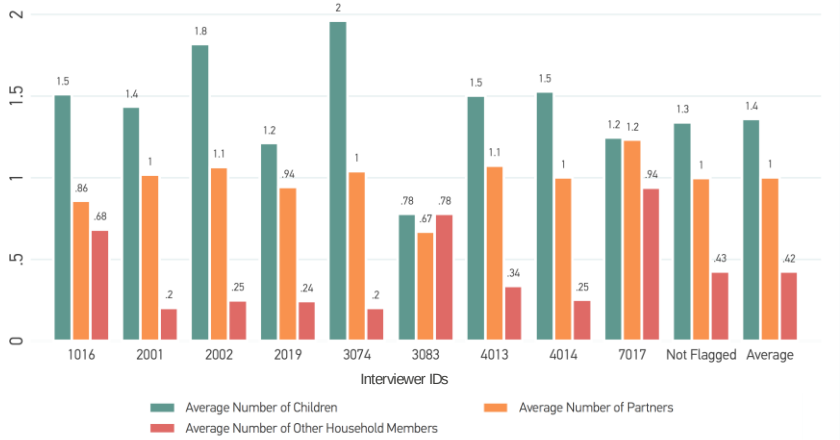
Out of 333 interviewers, for 17 we find an effect of the interview order on the number of children, for 16 we find an effect on the number of partners, and for 28 we

find an effect on the number of other household members.[8] The number of interviewers for which there is an effect on more than one variable is 7, who combined conducted 1,113 interviews (11.1% of the total). This is not to say that all these interviews were affected, but that a learning effect was detected for the interviewer that conducted these interviews. Our figures might suggest a more serious issue than in Ruckdeschel, Sauer, and Naderi (2016), as just three interviewers demonstrated learning behaviors in that analysis. However, to assess whether the impact is similar to that observed in Germany it is necessary to consider the size of the learning effect. In Germany, the removal of interviewers with a learning effect severely reduced the identified bias in reported births.

To see if this is the case in Belarus, we remove all the interviews carried out by the 7 anomalous interviewers and see if this has an impact on the three key variables. The results are presented in Figure 5. When we remove the anomalous interviewers the average number of children decreases from 1.41 to 1.38, the average number of partners decreases from 1.02 to 0.99, and the average number of other household members increases from 0.42 to 0.43. Looking at the signs, we find evidence of learning effects only for the number of other household members. We claim this is not strong evidence of manipulation, in contrast to the findings with regards to the GGS in Germany. We conclude that, given that interviewers are not randomly assigned to respondents, even a high number of interviewers for which the order of interview has a significant effect on the three key screening questions is not robust evidence of manipulation.

[8] 44.4% of these interviewers are inexperienced, 31.5% are experienced, and for the remaining 24.1% we don't know. The respective proportions in the sample are 28.2%, 25%, and 56.8%. The proportion of inexperienced interviewers among the anomalous ones is thus perceptibly higher than what we would have expected if the distribution among groups were random.

Figure 5: Average number of children, partners, and other household members for flagged interviewers



5.4 Interviewer characteristics

Testing Hypothesis 3 might lead us to a partial answer to this theoretical question. Looking at Table 1 and Table 2, we see that being a young student versus being a more experienced interviewer has a significant effect only on the number of partners. The size of the effect is small but not negligible; however, it is not clear evidence of young and inexperienced interviewers being more prone to fabrication. As Hypothesis 3 also predicts interview-order effects to be stronger for the young student group, we test this implication by adding the interaction between order of interview and interviewer's group into the model. We find no significant effects of the interaction term. This seems to suggest that the significant difference in the number of partners across the two groups may be due to non-random assignment of interviewers.

As a robustness check, we also run the model with additional controls as described in the methods section, which should capture most of the differences due to non-random assignment. Once we do this, the effects of both the interviewer's group dummy and the interaction term disappear.

5.5 Analysis of item-nonresponses

According to Hypothesis 4, the proportion of item-nonresponse should increase in later sections of the questionnaire. To test this hypothesis we construct a dataset where, for each question in the GGS, we have the number and the proportion of the three types of item-nonresponse allowed in the GGS: Non-Applicable, Refusal, and Don't Know. Since it can be difficult to attach a specific interpretation to the three types, especially because it is the interviewer who ultimately decides which type of nonresponse to record, we decided to look at them in aggregate. The results are presented in Figure A-2.

Looking at the graph, we do not see a clear pattern. There seem to be two parts of the questionnaire, one including sections 1 to 8 and the other the remaining sections. In the first part, we see no increase in item-nonresponse based on the order of sections. Section 6 appears to have the highest proportion of item-nonresponse in this part. This is quite expected, given that it is one of the most sensitive sections in the GGS and asks about contraceptive use, fertility intentions, and sexual health. In the second part we see a decreasing pattern and a perceptibly higher proportion of item-nonresponse compared to the first part. One aspect that should be noted is that the last three sections have a small number of questions (25, 19, and 11 respectively); thus they are more sensitive to the presence of questions with very high item-nonresponse, and they also include several questions about traditionally high nonresponse topics such as household income and occupational type. Overall, we see no clear evidence of a fatigue effect, and thus we reject Hypothesis 4.

We now analyze in more detail the three sections with the highest item-nonresponse. For section 9, we find that the two questions that push average item-nonresponse up are 2 and 29. Question 2 asks about the date when the respondent's partner entered their current activity status. Question 29 asks about the different types of income the respondent's partner has received in the last 12 months. We suggest that the reason for a high nonresponse rate for both questions is that respondents simply do not know the answer, and that it is not related to fatigue.

For section 10, item-nonresponse is driven upward by question 1, which asks the respondent to provide an estimate of the current market value of all real estate they own. This is clearly a difficult question, because it requires some knowledge of current house and land prices. As for section 9, we would say that the high nonresponse is due to an inability to answer, and not to fatigue.

Finally, for section 11, the questions responsible for the high item-nonresponse are 2 and 11. Question 2 asks about general fairness[9] while question 11 asks about attendance at religious services. Incorrect framing of the question might cause a high nonresponse

[9] "Do you think most people would try to take advantage of you if they had the chance, or would they try to be fair?"

to question 11. First, not all religions have institutionalised religious services and, second, this question is also asked of respondents that do not recognise themselves in any religious denomination. For question 2, the high nonresponse rate might be caused by the lack of exhaustive options (only "would take advantage" and "would try to be fair" are offered), although it is not clear why this happens for fairness and not for trust (question 1 in section 11).[10]

Overall, we see no evidence that the higher average item-nonresponse for the last three sections is caused by fatigue and we suspect that the nonresponse rates would have been the same for these sections even if they were not at the end of the survey.

6. Discussion

Comparing the TFR time-series obtained from the GGS with those provided by the Human Fertility Database and the UN World Population Prospects, we find no evidence of serious biases for the Generations and Gender Survey in Belarus. We obtain the same results when comparing the proportion of married women and childless women in the GGS with the same figures for Wave 6 of the WVS. In this case we notice that the GGS data seems to be more reliable and more stable over time compared to the WVS. This is also the case for the overall distribution of parity and marital status.

When we move to the analysis of interviewer and respondent effects, we find mixed evidence. We only find significant effects in the direction we would have expected if fabrication had occurred with regards to the household size. The number of children and partners shows no signs of being affected by fabrication. When assessing the impact of such effects on the household size and the overall quality of the data, we conclude that no or very weak distortion was introduced. We think that the most credible explanation for this apparent contradiction is that interviewers were not randomly assigned to respondents; therefore any regression model used to assess interviewer effects cannot be given a causal interpretation. Furthermore, because the contribution of each interviewer was rather small (an average of 30 responses), a few fraudulent interviewers would not compromise the overall data quality. It is entirely plausible that smaller households were harder to reach and thus were only contacted and interviewed later in the fieldwork window. This result should also make us reflect on the reliability of this analytical framework for detecting manipulation or fabrication (Schäfer et al. 2004; De Haas and Winker 2016). If we had looked exclusively at the analysis of interviewer and respondent effects we might have concluded that the data quality had been compromised. However,

[10] To confirm this intuition, we looked at Wave 6 WVS where the same question can be answered on a scale of 1 to 10, where 1 is "would take advantage" and 10 is "would try to be fair". Most respondents are indeed concentrated in the middle. This supports our conclusion.

as we have showed, GGS data does not seem to be affected by major distortions in its main indicators.

7. Conclusions

The main conclusion we draw from our analysis is that GGS Belarus 2017 has no serious distortions and that, overall, it is a reliable source of data both in terms of comparability with other sources and in terms of frequency and distribution of nonresponses.

From a practical point of view, we evaluate positively the new control procedures introduced in the GGS data-collection process. We believe that frequent reporting during fieldwork and issuing warnings to specific interviewers was effective in preventing extensive fabrication and consequent distortions. This is particularly relevant because in the GGS Belarus 2017 the interviewers were mostly young and inexperienced. This greatly reduced the cost of fieldwork and apparently did not harm the data quality.[11] If this model could be extended to other countries it would save a large amount of resources. The small distortions introduced, which we primarily attribute to non-random allocation of interviewers, could be better addressed by a greater degree of randomization in the allocation of interviews to interviewers, although this is hard to achieve in practice. This last finding is particularly relevant, because employing students is usually much cheaper than employing professional interviewers. If this cost reduction does not have a negative impact on the data quality, at least in the case of the GGS, then it might be worth assessing the advantages of pursuing this strategy.

The analysis does, however, have several limitations that should be noted. First, this analysis only looked at a specific type of data manipulation. Errors in survey data can originate from a multitude of sources and we did not seek to identify and measure these other forms. Such errors are laid out within the total survey error framework and further analysis should attempt to identify the broader range of errors that may exist within the Generations and Gender Survey and how these might be reduced and controlled. The resources to implement a total survey error framework are currently beyond those available to the Generations and Gender Survey but could be employed if sufficient investment in the survey is made.

Second, it is not possible to conclude definitively that the absence of data manipulation in Belarus was specifically due to the new fieldwork systems and controls in the Generations and Gender Survey. In the previous round of data collection, data manipulation was only identified in one country, and it could be that such manipulations

[11] We acknowledge that to reach this conclusion in a rigorous way a proper experiment would be needed. However, an experiment able to provide convincing evidence would need to be conducted on a large scale and would thus itself be extremely costly.

may reoccur in the collection of life histories in the new round of data collection. However, what is notable from the analysis is that the monitoring and identification of such issues is more feasible during the fieldwork than in previous rounds, allowing for corrective measures to be put in place and potentially reducing the impact of data manipulation on the overall data quality.

Finally, the fieldwork context in Belarus was specific and differed from the context in which previous manipulations have been identified. The manipulations in Germany occurred in a highly professional setting and it could be that this experience and the use of professional interviewers increased the incidence of such data manipulation. Therefore, it will be important to replicate this analysis in other countries planning to participate in the new round of the Generations and Gender Survey. This is especially true because a large proportion of countries in the new round of data collection will be collecting the data via a web survey. In a web environment the risk of errors in the collection of detailed life-history data remains, but the context is very different from that described here, and it will require further research to understand such errors and mitigate their effects on overall data quality.

8. Acknowledgements

The authors would like to acknowledge the extensive and professional work of the Belarussian Generations and Gender Survey National Team at the Belarussian State University. A special thanks is reserved for Professor David Rotman, Olga Tereschenko, and Victor Pravdivets. We also gratefully acknowledge the assistance of the United Nations Population Fund (UNFPA) and the coordination of Marianne Sakalova. We feel the results of this analysis give great credit to the professional and considerable expertise that the team brought to the project.

References

Aronson, J., Lustina, M.J., Good, C., Keough, K., Steele, C.M., and Brown, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology* 35(1): 29–46. doi:10.1006/jesp.1998.1371.

Becker, S., Feyisetan, K., and Makinwa-Adebusoye, P. (1995). The effect of the sex of interviewers on the quality of data in a Nigerian family planning questionnaire. *Studies in Family Planning* 26(4): 233–240. doi:10.2307/2137848.

Benstead, L.J. (2013). Effects of interviewer: Respondent gender interaction on attitudes toward women and politics: Findings from Morocco. *International Journal of Public Opinion Research* 26(3): 369–383. doi:10.1093/ijpor/edt024.

Blaydes, L. and Gillum, R.M. (2013). Religiosity-of-interviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt. *Politics and Religion* 6(3): 459–482. doi:10.1017/S1755048312000557.

Catania, J.A., Binson, D., Canchola, J., Pollack, L.M., Hauck, W., and Coates, T.J. (1996). Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly* 60(3): 345–375. doi:10.1086/297758.

Davis, D.W. and Silver, B.D. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science* 47(1): 33–45. doi:10.1111/1540-5907.00003.

De Haas, S. and Winker, P. (2016). Detecting fraudulent interviewers by improved clustering methods: The case of falsifications of answers to parts of a questionnaire. *Journal of Official Statistics* 32(3): 643–660. doi:10.1515/jos-2016-0033.

Dykema, J., Diloreto, K., Price, J.L., White, E., and Schaeffer, N.C. (2012). ACASI gender-of-interviewer voice effects on reports to questions about sensitive behaviors among young adults. *Public Opinion Quarterly* 76(2): 311–325. doi:10.1093/poq/nfs021.

Fadel, L., Emery,T., and Gauthier, A.H. (2020). Current and future contributions of the Generations and Gender Programme to lifecourse research. In: Falkingham, J., Evandrou, M., and Vlachantoni, A. (eds.). *Handbook on demographic change and the lifecourse*. Cheltenham: Edward Elgar Publishing: 57–68. doi:10.4337/9781788974875.00012.

Flores-Macias, F. and Lawson, C. (2008). Effects of interviewer gender on survey responses: Findings from a household survey in Mexico. *International Journal of Public Opinion Research* 20(1): 100–110. doi:10.1093/ijpor/edn007.

Gallagher, A.M. and De Lisi, R. (1994). Gender differences in scholastic aptitude test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology* 86(2): 204–211. doi:10.1037/0022-0663.86.2.204.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen B. (2014). World values survey: round six: Country-pooled datafile version. JD Systems Institute. http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp.

Kane, E.W. and Macaulay, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly* 57(1): 1–28. doi:10.1086/269352.

Klüsener, S., Jasilioniene, A., and Yuodeshko, V. (2019). Retraditionalization as a pathway to escape lowest-low fertility? Characteristics and prospects of the Eastern European "baby boom". Rostock: Max Planck Institute for Demographic Research (No. WP-2019-014). doi:10.4054/MPIDR-WP-2019-014.

Koczela, S., Furlong, C., McCarthy, J., and Mushtaq, A. (2015). Curbstoning and beyond: Confronting data fabrication in survey research. *Statistical Journal of the IAOS* 31(3): 413–422. doi:10.3233/SJI-150917.

Kreyenfeld, M., Hornung, A., Kubisch, K., and Jaschinski, I. (2010). Fertility and union histories from German GGS data: Some critical reflections. Rostock: Max Planck Institute for Demographic Research Working Paper 23. doi:10.4054/MPIDR-WP-2010-023.

Lau, C.Q. (2018). The influence of interviewer characteristics on support for democracy and political engagement in Sub-Saharan Africa. *International Journal of Social Research Methodology* 21(4): 467–486. doi:10.1080/13645579.2017.1407087.

Lipps, O. and Pollien, A. (2010). Effects of interviewer experience on components of nonresponse in the European Social Survey. *Field Methods* 23(2): 156–172. doi:10.1177/1525822X10387770.

MPIDR and Vienna Institute of Demography. 2018. Human Fertility Database.

Olson, K. and Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly* 71(2): 273–286. doi:10.1093/poq/nfm007.

Polglase, G. (2013). Higher education as soft power in the Eastern Partnership: The case of Belarus. *Eastern Journal of European Studies* 4(2): 111.

Ruckdeschel, K., Sauer, L., and Naderi, R. (2016). Reliability of retrospective event histories within the German Generations and Gender Survey: The role of interviewer and survey design factors. *Demographic Research* 34(11): 321–358. doi:10.4054/DemRes.2016.34.11.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., and Wagner, G.G. (2004). Automatic identification of faked and fraudulent interviews in surveys by two different methods. (DIW Discussion Papers 441). Berlin: DIW.

Shchurko, T. (2012). "Compulsory motherhood": The female reproductive body as regulated by the state (Based on the Analysis of Newspaper Sovetskaia Belorussiia). *Laboratorium: Russian Review of Social Research* 4(2): 69–90.

Shchurko, T. (2017). 'Gender education' in the post-Soviet Belarus: Between authoritarian power, neoliberal ideology, and democratic institutions. *Policy Futures in Education* 16(4): 434–448. doi:10.1177/1478210317719779.

Tikhonova, L.E. (2004). Marriage and family relations in the Republic of Belarus. *Sociological Research* 43(1): 83–91. doi:10.1080/10610154.2004.11068579.

Tu, S.-H. and Liao, P.-S. (2007). Social distance, respondent cooperation and item nonresponse in sex survey. *Quality and Quantity* 41(2): 177–199. doi:10.1007/s11135-007-9088-0.

United Nations (2017). *World population prospects: The 2017 revision*. New York: United Nations.

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.

World Bank and OECD (2017). World Bank national accounts data, and OECD National Accounts data files. https://data.worldbank.org/country/belarus.

Appendix 1

Figure A-1 shows the distribution of marital status in Wave 6 of the WVS and the GGS Belarus 2017. The two distributions are fairly similar. In the GGS there is a higher proportion of married and cohabiting individuals and a lower proportion of widowed individuals.

Figure A-1: Marital status in Belarus from the WVS (2011) and GGS (2017)

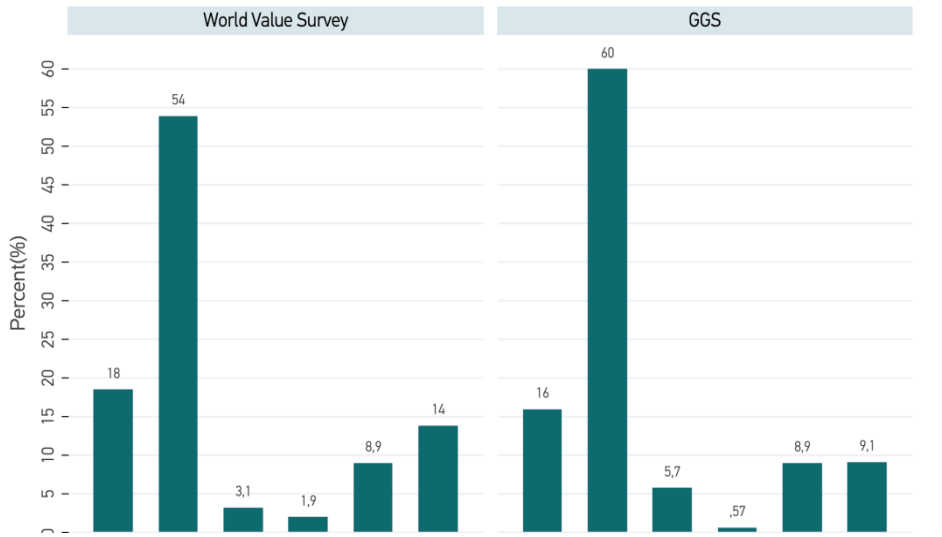
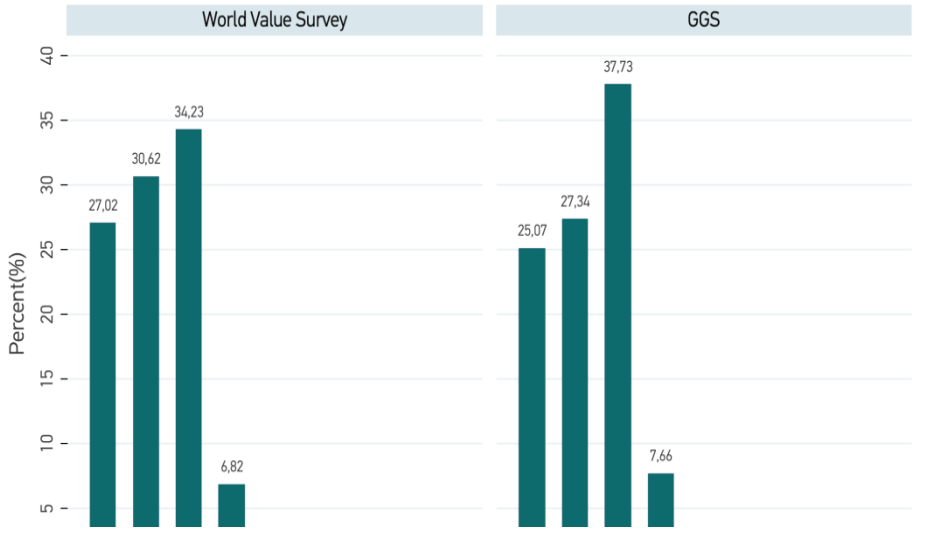


Figure A-2: Respondents' number of children in Belarus from the WVS (2011) and GGS (2017)



Appendix 2

This appendix describes in more detail the complex multilevel model we mentioned in Section 5. Exploiting the large amount of information on interviewers and respondents available in GGS Belarus 2017, we are able to add more controls than in the basic model. Based on the literature on interviewer effects and on the hypotheses we presented in Section 4, we decided to include the following variables:

- Order of the interview, as the main explanatory variable.
- Total number of interviews conducted by the interviewer.
- Respondent's age, income, sex, education, activity status, marital status, education, religion, and region of origin.
- Interviewer's age, sex, education, and religion.
- A three-category variable for age distance.
- A binary variable for same region of origin.
- Two-way interaction between respondent's and interviewer's religion.
- Three-way interaction between respondent's sex, interviewer's sex, and age distance.
- Four-way interaction between respondent's sex, interviewer's sex, respondent's religion, and respondent's education.

As in the basic model, we adjust standard errors to account for interviewer clustering. Some clarification of the content and coding of some variables is in order.

The activity status variable included in this model has 4 categories instead of the 12 present in the corresponding GGS variable. We decided to recode the original variable so as not to have too small subsamples. We chose four categories with the aim of creating homogenous groups: 'Currently Working', 'Temporarily Not Working', 'Permanently Not Working', and 'Studying'. We recoded 'Employed', 'Self-Employed', 'Helping Family Member on a Family Farm', 'In Military or Civic Service', and 'Homemaker' as 'Employed'. We recoded 'Unemployed', 'On Maternity Leave', and 'On Parental Leave or Childcare Leave' as 'Temporarily Not Working'. We recoded 'Retired' and 'Ill or Disabled for a Long Time or Permanently' as 'Permanently not Working'. Finally, we recoded 'Student, in School, Vocational Training' as 'Studying'. Those respondents who indicated 'Other' were coded as missing, since we had no information that could help us sort them into one of the four categories and their number was too small to leave them in a separate category.

The education variable for both interviewers and respondents was recoded from 8 categories to just 2: 'Below University' and 'University or Above'. We did this to avoid the sample size being too small, especially because we also use education level in the

interaction terms. We recoded 'Early Childhood Education', 'Primary Education', 'Lower Secondary Education', 'Upper Secondary Education', and 'Post-Secondary Non-Tertiary Education' as 'Below University'. We recoded 'Bachelor or Equivalent', 'Master or Equivalent', and 'Doctoral or Equivalent' as 'University or Above'.

Finally, the religion variable for both interviewers and respondents was recoded from 11 categories to 3. We did this both to avoid having a too small sample size for subgroups and because more than 80% of the sample declared themselves to be orthodox, so the remaining groups were very small. We recoded 'Orthodox (e.g., Greek or Russian)' as 'Orthodox', all other religious groups as 'Non-Orthodox', and those who chose 'None' an 'Non-Religious'.

Appendix 3

To assess the reliability of household size distribution in the GGS Belarus 2017, we compared it to that in the 2009 Census. We conducted this analysis at the regional level, since this is the finest geographical disaggregation that the GGS data allows. We present the results in the following figure.

Figure A-3: Over- and under-representation of household size in the GGS, by region