Break offs in web surveys of the Generations and Gender Programme: Patterns and predictors

Dr. Arieke Rijken (GGP fieldwork coordinator, rijken@nidi.nl)
Netherlands Interdisciplinary Demographic Institute
Extended abstract submitted to EPC 2022
Theme: Data and Methods

Introduction and Hypotheses

In 2020, the Gender and Generations Programme (GGP), a data infrastructure for research on population and family dynamics, started a new round of data collection with fresh samples in all participating countries: GGP II (or sometimes as GGP 2020, although the fieldwork will spread over several years). Whereas previous Generations and Gender Surveys (GGS) were all face-to-face, in this new round of data collection many countries decided to conduct their GGS as a web-only survey. This trend started before the COVID-19 Pandemic, driven by increased costs of face-to face interviews, but was accelerated by lockdown measures due to the COVID Pandemic.

A major challenge in conducting web surveys are break offs (e.g. Mittereder & West 2021; Peytchev 2009), resulting in partially completed questionnaires, which leads to smaller analytical samples and potentially to breakoff bias. Since the GGS has a long questionnaire for web survey standards (about 45 minutes) and the length is affected by respondent characteristics that are very central to the GGP (such as number of children), the problem of break offs requires special attention. Findings of this study may be used for the design of future GGP questionnaires.

I use recent data from GGS web surveys in several countries to study break offs. I present descriptive results on cross-national variation in the percentage of break offs and in break off patterns: where in the questionnaire do respondents break off?

Furthermore, I test hypotheses on the association between respondent characteristic and breaking off and test them using logistic regression analysis. The length of the questionnaire is heavily affected by the life history and family composition of the respondent due to routing and loops. For example, when the respondent has a partner, many questions are asked on the partner's background characteristics, the partner work situation, and on the relationship itself are asked (e.g. on division of tasks). Also, respondents are asked to answer a set of question about each of the ex-partners they lived with, and another set of questions about each of their children (biological, adopted and stepchildren).

Therefore, I hypothesize that respondents are less likely to complete the questionnaire when they have a partner, the more-ex partners they lived with, and the more children they have. Several arguments can be given for these hypotheses. First of all, each of these factors increase the length of the questionnaire. Second, apart from the length, respondents may find the questionnaire more tedious when they repeatedly have to answer the same set of questions for multiple ex-partners and for multiple children. Third, respondents may feel they are not entitled to provide detailed information about other people than themselves. However, one could also argue that respondents with a partner or with children find the survey more suited to them, or more interesting, which may counteract the effects described above.

Data and Method

Data

I will use all GSS web surveys of GGP II, wave 1 that are available. For the provisional analyses presented in this extended abstract I included data from GGS Norway (2020), GGS Denmark (2021),

from GGS pilots for the first wave in Estonia (2021) and Hong Kong (2021) and from currently ongoing GGS fieldwork in Czechia. These surveys were all web-only; no alternative mode of data collection was offered, except in Czechia, where respondents can choose between CAPI and CAWI (only the CAWI interviews are used in this paper). Before the conference I will be able to add data from more countries and full first waves (instead of only pilot data) from Estonia and probably also from Hong Kong.

For all surveys (including the two pilots) random samples were drawn, either from population registers (Norway, Denmark, Estonia) or from address or building registers (Hong Kong and Czech Republic). In the latter situation, the last birthday method was used to select a respondent from the household. The age range varied between countries: Denmark: 18-49, Norway: 18-54, Estonia and Hong Kong: 18- 59, and Czechia: 18-69. Fieldwork was conducted either by National Statistical Agencies (Norway and Denmark), by commercial survey agencies (Czech Republic and Estonia) or by the university where the national scientific team is located (Hong Kong). Respondents were invited to participate via e-mail (Norway, Denmark, Estonia), ordinary mail (Hong Kong) or at their door by employees of the survey agency (Czechia). In all cases the fieldwork was prepared and conducted in close collaboration between the GGP central coordination team, using a centrally coded questionnaire.

Analytical strategy and variables

First I will present descriptive analyses by country, showing which percentage of respondents who start the questionnaire also complete it where in the questionnaire respondents break off. In these analyses I included all respondents who logged in to the survey, including those who did not answer any question after logging in. The sample sizes are as follows: Denmark N = 10,266, Norway: N = 8,026, Estonia (pilot): N = 3.424 (pilot) Hong Kong (pilot) , N = 1,728 pilot) and Czechia (fieldwork ongoing): N = 616.

Next, I will present a logistic regression analysis in order to test the hypotheses on the association between respondent characteristics on survey completion. For this analysis I pooled the data of all countries and included country dummies. I had to use survey answers to measure the respondent characteristics (the independent variables in the analysis), therefore I could not include all respondents who logged in, they had to have answered at least the questions needed to measure the independent variables (N = 18,054 for the model including all predictors).

The dependent variable in the logistic regression analysis is whether the respondent completed the questionnaire (1) or not (0). I regarded the questionnaire as completed if the last question of the 'standard questionnaire' is answered (when the respondent answered 'Don't Know' or 'Refusal' the questionnaire is also considered completed). Some countries added a country-specific module at the end of the questionnaire, these modules are disregarded in this paper. The independent variables are sex (1 = female, 0 = male), age group (18-29;30-49 and 50-70), educational level (low; medium; high) partner status (has partner = 1, no partner = 0), number of children (0; 1; 2; 3+), including biological children, adopted children and stepchildren and number of ex-partners with whom the respondents has lived in one household (0; 1; 2; 3+).

Preliminary Results

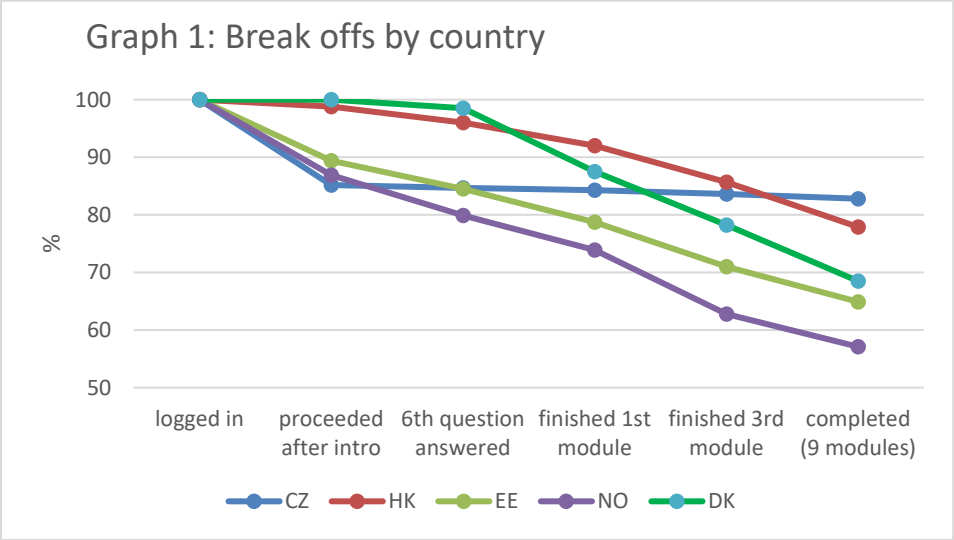
Descriptive results: break off patterns

Graph 1 shows descriptive r the break off pattern by country. The graphs is based on all respondents who logged in to the survey (= 100%) and shows which percentage of these respondents reached certain points in the questionnaire; which percentage…

- clicked on 'proceed' after the introduction text on the first page of the survey;

- answered the 6th question of the questionnaire (the first questions are about background characteristics such sex, year and month of birth, country of birth and education);
- finished the first module of the questionnaire;
- finished the third module of the questionnaire;[1]
- completed the ninth module, that is, completed the standard questionnaire.

First of all, the graph shows that the percentage of respondents breaking off varies a lot by country: from17% in Czechia to 43% in Norway. If we define 'proceeding after the introduction text' as actually starting the survey (instead of logging in), the percentage of break offs is lower, especially in Norway, Estonia and Czechia. In Norway and Estonia there was no landing page, respondents received an individualized link containing their access code in an e-mail, and thus logged in more easily than in other countries, by just clicking on that link. The graph also shows that respondents who finish the first three modules of the questionnaire, have a high likelihood of finishing the remaining six modules as well; most break offs occur relatively early in the questionnaire. In the paper, these results will be further discussed while paying attention to the content of the questionnaire and differences between countries in fieldwork procedures and incentives.



Predictors of questionnaire completion

Table 1 shows the results of a logistic regression analysis estimating the effects of certain respondent characteristics on questionnaire completion. Model 1 only *Female*, *Age group*, and *Has partner* as predictors. These variables are based on questions from the first half of the first module questionnaire (about 20 questions). Model 2 also contains the number of ex-partners and number of children; we only have this information for respondents who finished at least half of the second module of the questionnaire (therefore Model 2 is based on a smaller sample than Model 1).

Model 1 shows that women are more likely to complete the questionnaire than men, that respondents in the age group 30-49 are less likely to complete the questionnaire than younger respondents (18-29), whereas there is no difference between the 50+ age group and the 18-29 age

[1] The first three modules are the most important ones in the questionnaire, containing questions on general demographics, partnership and fertility histories and fertility intentions and reproductive health.

group. Finally, the OR of *Has partner* shows that respondents with a partner are less likely to complete the questionnaire than respondents without a partner, which is in line with my hypothesis.

Model 2 shows that the more children and the more ex-partners a respondent has, the lower the chances he or she will complete the questionnaire, which is also in line with my hypotheses. However, adding these variables changes the effect of having a partner: in Model 2 having a partner has a positive effect on questionnaire completion, contradicting my hypothesis. This suggests that the effect of having a partner on questionnaire completion might be mediated by the number of children. Also, the effect of age has changed in Model 2 compared to Model 1, with the 50+ age group now being most likely to finish the questionnaire. Perhaps this positive effect only appears in Model 2 because the 50+ are more likely to have multiple ex-partners, which is now taken into account. In the paper I will look into this more carefully. I may also add more predictors, differentiate between biological and stepchildren and I aim to control for device (preliminary results from Norway and Czechia show that respondents filling out the survey on a phone are less likely to complete the questionnaire).

In short, these results show that there is indeed break off bias. Representativity analyses from each country are needed to say more about the severity of this bias (combined with non-response bias).

Table 1: Logistic regression models of predictors of questionnaire completion

		Model 1 (N = 19,580)	Model 2 (N = 18,054)
		OR	OR
Constant		43.66***	92,33***
Female		1.27***	1,20***
Age group, ref:	18-29		
	30-49	0.76***	0,94
	50-79	1.00	1,27***
Education, ref = low			
	middle	1.25**	1,19***
	high	1.78***	1,75**
Has partner		0.82***	1,19***
Children, ref:	0		
	1		0,87
	2		0,78***
	3+		0,72***
Ex-partners, ref:	0		
	1		0,77***
	2		0,67***
	3+		0,35***

Note: Data from NO, DK, EE, CZ, and HK are pooled.
Models include country dummies (not shown).

References:

Mittereder, F. and West, B. T. (2021). A dyntamic survival modeling approach to the prediction of web survey breakoff., *Journal of Survey Statistics and Methodology*, smab015, https://doi.org/10.1093/jssam/smab015

Peytchev, A. (2009). Survey breakoff, *Public Opinion Quarterly*, 73(1), 74–97, https://doi.org/10.1093/poq/nfp014