Generations and Gender Programme Preparatory Phase Project (GGP-5D)

Laurent Toulemon, Géraldine Charrance, Paul Cochet, Ignacio Pardo, Arieke Rijken

Generations and Gender Programme Preparatory Phase Project (GGP-5D)

Technical paper on analysing multi-mode GGS data for users

Work package 1: **Technical Design**
Grant Agreement Number: **101079357**
Project acronym: **GGP-5D**
Project full title: **The Generations and Gender Programme Preparatory Phase Project**

Due delivery date: **31 December 2023**
Actual delivery date: **21 December 2023**
Organization name of lead participant for this deliverable**: French Institute for Demographic Studies (INED)**

Dissemination level: **Public**

Funded by the European Union

Document Control Sheet

Deliverable number:	D1.1
Deliverable responsible:	Laurent Toulemon
Work package:	1. Technical Design
Editor(s):	Vytenis Deimantas

Author (s)		
Name	***Organization***	***E-mail***
Laurent Toulemon	***INED***	***toulemon@ined.fr***
Géraldine Charrance	***INED***	***geraldine.charrance@ined.fr***
Paul Cochet	***INED***	***paul.cochet@ined.fr***
Ignacio Pardo	***Universidad de la República***	***ignacio.pardo@cienciassociales.edu.uy***
Arieke Rijken	***NIDI***	***Rijken@nidi.nl***

Document Revision History			
Version	***Date***	***Modifications Introduced***	
		Modification reason	***Modified by***
V1			
V2			
V3			

Executive summary

This technical guide presents the State of the art regarding multi-mode data collection, related to the Generations and Gender Surveys (GGS – Round II). Mode effects may come from selection of the respondents (mode selection effects) or from the way respondents understand the questions and answer (mode measurement effects).
Mode effects are estimated using GGS Round II data, using the French GGS test and the Uruguayan GGS, which both have collected data on two modes, from different representative subsamples; internet and telephone in France, Internet and face-to-face in Uruguay. Three main data quality indices are computed and used.
In both countries, non-response are more frequent in the internet questionnaires than in the questionnaires filled in with an interviewer; straight-lining, the tendency for some respondents to give always the same answer when facing a battery of similar questions, is also more frequent on internet.
Finally, questionnaires filled in on Internet are more diverse, with less frequent responses. This is likely due to lower social desirability bias, as the respondents are not interacting with an interviewer.
Based on these comparisons, the positive global assessment of the push-to-web strategy conducted by the GGP is confirmed. Nevertheless, new methods are arising and new methodologies are tested. Mode effects are one among many methodological challenges, and the ongoing methodological efforts of the GGS central hub and national teams must be pursued.

Table of Contents

Definitions and acronyms

CAPI: Computer assisted personal interview
CATI: Computer assisted telephone interview
CAWI: Computer assisted web interview (in this instance self-administered)
F2F: face-to-face (interview)
GGP: Generations and Gender Programme
GGS: Generations and Gender Surveys
GGS-Round II: Generations and Gender Surveys, second round collected in the 2020s
PAPI: Paper and Pencil interview (in this instance self-administered)

1 Mode effects and the GGP

1.1 Data collection mode by country

The Round II of the Generations and Gender Surveys (GGS) moved to Computer Assisted Web Interviewing (CAWI), based on the analysis that Push-to-Web was the most cost-efficient way to collect data for the GGS – Round II set of surveys. Push-to-Web means contacting potential respondents and asking them to fill in the GGS questionnaire on the internet, from a personal computer, a tablet or a smartphone. A pilot study was conducted in 2018 in Croatia, Germany and Portugal, comparing Face-to-face and Push-to-Web designs. The first conclusion of that pilot was that the push-to-web design was equally successful than the face-to-face design in terms of response rate and non-response bias in two countries, Germany and Croatia. In Portugal, the push-to-web design was not successful, due to the lack of a sampling frame (list of potential respondents), and the low internet coverage. The second conclusion of the Pilot was that a mix of unconditional and conditional incentives worked best to increase response rates in CAWI. Overall, the pilot allowed concluding that it was possible to use a push-to-web design as long as a sampling frame of individuals was available (Lugtig et al., 2022).

After this positive experiment and as the data collection cost is lower than with other modes, CAWI became the main mode of data collection. Nevertheless, the previous round had taken place through Computer Assisted Personal Interviews (CAPI), with interviewers meeting respondents face-to-face, the only exception being Hungary, with a partial data collection in CAWI for its 2012 wave (questionnaires partly filled in through CAWI, and completed in CAPI). Furthermore, the very first surveys of the GGP – Round II series conducted in Belarus, Kazakhstan and Latvia were also conducted in CAPI. The choice of data collection mode was thus not restricted to CAWI for the GGP – Round II series of surveys, but CAWI was defined as the preferred data collection mode, eventually completed by another mode.

Table 1 presents the distribution of countries by data collection mode. Most countries chose CAWI as the only data collection mode. Czechia offered CAWI and CAPI, but the latter on a very small scale. Germany and Sweden complemented CAWI with PAPI (self-administered questionnaire sent by postal mail). The Czech Republic and Uruguay organized the data collection mainly through CAPI, with a methodological sample based on CAWI. France will complement CAWI by telephone interviews (CATI), after a two-week CAWI only data collection.

Table 1. An overview of GGS-II baseline data resource and coverage

	Net sample size	Age range	Resp-onse rate	Mode	Data collection
					Baseline (wave 1)
Argentina	2,433	18-79	41%	CAPI	Aug – Dec 2022
Austria	8,265	18-59	38%	CAWI	Oct 2022 – Mar 2023
Belarus	9,994	18-79	76%	CAPI	Apr – Nov 2017
Croatia	6,000*	18-54	30%*	CAWI	May – Jun 2023
Czech Republic[a]	5,583	18-69	17%	CAWI(98%), CAPI(2%)	Oct 2020 – Jul 2022
Denmark	8,269	18-49	20%	CAWI	Mar – Jun 2021
Estonia[a]	8,992	18-59	29%	CAWI	Oct 2021 – Feb 2022
Finland	3,388	18-54	17%	CAWI	Oct 2021 – Mar 2022
France[a]	10,000*	18-79	34%*	CAWI, CATI	In preparation
Germany	22,281	18-59	21%	CAWI (85%), PAPI (15%)	Jun 2021 – Feb 2022
Hong Kong SAR[a]	5,088	18-59	32%	CAWI	Feb – Mar 2023
Italy	10,000*	18-59	NA	CAWI, CAPI	In preparation
Kazakhstan	14,857	18-79	93%	CAPI	Apr – Oct 2018
Latvia	2,298	18-79	57%	CAPI	Sep – Nov 2018
Moldova	10,044	15-79	50%	CAPI	Jan – Dec 2020
Netherlands	7,000*	18-59	29%*	CAWI	Nov 2022 – July2023
Norway	5,031	18-54	33%	CAWI	Nov – Dec 2020
Sweden	8,082	18-59	27%	CAWI(67%), PAPI (33%)	Mar – Aug 2021
Uruguay	7,245	18-79	42%	CAPI (86%), CAWI (14%	Oct 2021 – Oct 2022
United Kingdom	7,723	18-59	15%	CAWI	Sept 2022 – March 2023

Notes:
1. Net sample size include respondents who completed at least the first two modules (Demographics and Life Histories);
2. Taiwan has conducted a pilot study in CAWI mode in January 2023.
3. * estimated number. Croatia and the Netherlands are in field in the mid 2023. The net sample sizes and response rates in these two countries are estimated based on the data that have been collected so far. France and Italy are preparing for fieldwork. The net sample sizes are targeted sample size based on prior knowledge. The response rate in France is estimated based on the French pilot study in 2022.
4. [a]. A Pilot survey has been conducted before conducting the baseline wave.

1.2 What are mode effects?

Describing the life cycle of surveys in a design perspective, Groves et al. (2004, Chapter 2) identify representation and measurement in the elaboration of survey statistics. Representation refers to defining and sampling the target population, getting answers and eventually adjusting the sample of respondents to the original population. Measurement relates to constructing the questionnaire, collecting and editing the answers.

Data collection mode can impact both dimensions (Klausch, 2014; Hox et al., 2015). ***Mode selection effects*** describe the impact of data collection mode on the representation aspect, mainly through selection of potential respondents. Among people selected to answer, some cannot be reached through all modes: for instance people without an internet connection cannot easily answer a CAWI interview; respondents can be specific, and some adjustment can be performed is their characteristics are observable and actually observed in the sampling frame or in an external dataset. ***Mode measurement effects*** refer to the same respondents giving different answers, depending on the data collection mode. In practice, these mode effects cannot be observed directly, as the same respondents are not

interviewed simultaneously through two modes (except is specific experiments, see e.g. Biemer, 2001), but measurement mode effects can be precisely observed when the mode selection effect is negligible or controlled for: different distributions of answers coming from two identical sub-samples (e.g. two representative samples of the same population) are interpreted as the consequence of mode measurement effects. These effects and their measurement will be presented in part 2 and tested in Part 3 of this deliverable.

1.3 Mode effects and data quality issues

Mode effects are one among many issues related to data quality. When the data collection process includes more than one mode of data collection, and when the two subsamples lead to different estimates, this raises the question of adjusting the procedure in order to get consistent results. But when using one mode of data collection, the consistency of the results is due to the unicity of the data collection, and does not guarantee that no distortion is present. As a matter of fact, selection effects may also be present, with a selection into the sample, as far as the sampling frame is not based on a representative sample, or if the response rate is lower than 100%; similarly, each survey may be subject to measurement bias, even if it is based on one data collection mode only.

Thus, multi-mode data collection can be viewed as a very efficient way to identify such effects, more than a problem that should be fixed. Furthermore, using multiple mode can be efficient to improve data quality, especially if the subsamples reached through each modes are different and complementary. Post-stratification of the sample is possible when external information is available, and allows correcting for selection effects. Similarly, the information on the mode of data collection for each sub-sample can enrich the analyses with considering measurement effects, as far as it is well documented.

2 Defining and measuring mode effects

2.1 Mode selection effects

Mode selection effects are due to the composition of the sample of respondents: respondents who answer in one mode are not the same than those who answer in another. Mode selection effects can occur because one of the modes is not accessible (digital illiteracy, lack of equipment) or because of personal preferences. It is sometimes possible to identify the factors that influence the propensity to answer the survey, and therefore the composition of the sample, with using auxiliary information available in the sampling frame. This is known as ignorable non-response; the selection in the sample is MAR, missing at random. When these mechanisms are not observable, non-response cannot be ignored without introducing bias; this is called non-ignorable non-response, leading to MNAR, missing not at random (Rubin, 1976).
If, conditional to observable characteristics, non-respondents are missing at random, an adjustment on observables, most often by weighting procedures based on a post-stratification on known margins, is efficient to control for selection into the sample. If the selection into the sample is not at random (or related to unobservable characteristics), low response rates may lead to bias in the survey statistics. For instance, if, everything equal, some people being temporarily upset answer more or less to the survey than the others, the sample results will refer to a subgroup of the population with more or fewer people upset. It could be that people temporarily upset are keen to answer internet self-administered surveys, and not to answer face-to-face. The risk of unobservable selection effects is larger when the overall response rates are low.
The risk of selection effects can be measured with using R-Indicators, which measure the proximity of the sample structure with a representative sample. It is based on the standard deviation of inclusion probabilities. When the structure of the original sample is known, the inclusion probabilities can be estimated for each stratum, and partial R-indicators measure the standard deviation of inclusion probabilities related to one variable, within subgroups formed based on other variables (Schouten et al., 2012). The same logic is used when each individual response is weighted, the weighted sample being made similar to the original representative sample (or the whole population) through post-stratification methods.

NMAR selection occurs when the probability to answer the survey is related to an interest for the theme of the survey (Groves et al. 2000). In such case, the correlation between inclusion on the sample and a variable under study can be explicitly modelled, like in Heckman (1979) selection hypothesis and procedures. We will see in section 2.3 that such models may lead to large adjustments in the final estimates.

As response rates in surveys are declining, there are growing concerns about selection bias. Using several data collection modes may lead to mose selection effects, but it may also increase the overall response rates and limit heterogeneity in response rates. When subsamples coming from different data collection modes are complementary, the overall representativeness of the sample can be improved by merging the two subsamples (Castell et al. 2023b).

2.2 Mode measurement effects

Answering a survey implies some effort. Survey methodology includes methods to make answers as homogenous as possible: every respondent must understand and answer the same question. Pilot surveys and focus groups allows testing questionnaires and checking the relevance and accuracy of wording of questions and answers. Some questions can be more difficult to handle through one data collection mode, leading to specific mode measurement effects. When an interviewer is asking the questions, there is a risk that the interaction influences the answers (interviewer effect). On the other hand the interviewer is supposed to remain neutral, but also to explain the questions or help the respondent to think about her or his answer.

The concept of "satisficing" groups many answering strategies leading to present a "good image" or to limit the burden of answering (e.g. with answering rapidly or systematically). According to Krosnick (1991), satisficing can manifest in:

- choosing explicitly offered no-opinion or 'don't know' response option
- choosing socially desirable responses
- non-differentiation or straight-lining when a battery of questions asks for ratings of multiple objects on the same response scale
- acquiescence response bias, which is the tendency to agree with any assertion, regardless of its content
- selecting the first reasonable looking option
- randomly selecting a response
- skipping items
- abandoning the survey or terminating the survey early
- rushing on online surveys
- choosing minimally acceptable answers when verbal answers are required

We will focus here on the three first items, which we can identify and measure through multivariate regression on individual GGS data. We will chose sets of variables in order to identify large mode effects, without focusing on one or the other topic. We will compare CAWI questionnaires with alternative modes based on interviewers, CATI by telephone or CAPI from face-to-face interviews.

Partial non-response is easy to measure on a subset of the variables. The lower the proportion of questions without an answer, the better. Non-response are supposed to be more frequent on CAWI, because no interviewer is here to insist for getting an answer if the respondent hesitates or is embarrassed by a question. We will disentangle non-response to quantitative questions (related to frequency, numbers, quantities) and qualitative questions with a fixed number of proposed answers, as partial non response (and mode measurement effects) are more likely for quantitative questions.

Diversity of answers can be estimated by the proportion of respondents offering "rare answers" to the survey. ***Social desirability*** is likely to lead respondents to prefer giving answers that they perceive as normative. Answers chosen by most people are likely to be identified as normative, so that respondents subject to social desirability bias are more likely to give frequent answers. The more diverse the answers (and thus the more frequent the "rare answers"), the better. Social desirability is more likely when an interviewer is asking the questions, so that the diversity of answers is supposed to be larger on CAWI.

Variance in the answers to a battery of questions. When a set of questions asks for ratings of multiple objects on the same response scale, ***straight-lining*** describes situations when a respondent gives the same answers to all the questions of the battery. The larger the variance, the better. In the absence of an interviewer, respondents by Internet are more likely to answer rapidly to such batteries, so that straight-lining is supposed to be more frequent on CAWI. The similarity of answers is larger when the consistency of responses corresponds to identical modalities in the questions of the set. This is the case in some batteries in the GGS questionnaire, while for others an overall consistency corresponds to different modalities. Straight-lining is easier to identify in the latter case, while in the former a large similarity of answers is more likely to occur, with smaller mode effects.

These mode measurement effects can be identified with comparing similar subsamples, without mode selection effects. Mode selection effects are less likely when random subsamples are compared (or when the assignment to one or the other response mode is random), allowing assuming no selection effect or the same selection effects. This is the case in experimental designs or studies (such as the GGP 3 country pilot). In all instances, the comparisons can be performed on standardized samples (with post-stratification on the same margins), as well as with using standardization hope to multivariate analyses or matching the subsamples by a propensity score to answer through one mode or the other (Rosenbaum and Rubin, 1983) in order to make the two subsamples comparable.

2.3 Two examples of selection and measurement effects

Several mode effects can thus be identified. Mode selection effects can be easily controlled if selection is occurring at random; they are more difficult to handle if selection is directly related to the topic of interest; mode measurement effects are due to the interaction between respondents and interviewers, or between respondents and the device on which they are filling the questionnaire.

The same team from the French Institute of Statistics (INSEE) produced two recent publications showing that no general rule can apply.
The Epicov survey is a panel survey conducted just after the first lockdown, in order to estimate COVID-19 prevalence and COVID-19-related changes in behaviour (Warszawski et al. 2022). A large sample was selected: 370 000 adults were identified to answer an Internet survey in 2020; a subsample was interviewed by telephone and internet. The overall response rates is 37%, significantly higher among the multimode subsamples (46% = 18% CATI + 28% CAWI) than for single CAWI subsamples (35%). Using sophisticated econometric methods, the authors try and identify mode selection and mode measurement effect. They found the mode measurement effects to be limited but, regarding COVID-19 prevalence, the found large differences between subsamples: 23,5% of respondents in the multimode subsamples declared at least one COVID-19-related symptom, as against 27,2% in the monomode subsamples. Using a Heckman election model, they estimate the inclusion probability and find a very high correlation of residuals (0.5) between prevalence and participation. They show that mode measurement effects are limited, and cannot explain more than 30% of the discrepancy at most. This very large selection effect lead the authors to propose a final estimate of 16% for the overall prevalence (Castell et al. 2023a). Considering the sample size (130 000 respondents), the estimated variance of these estimates is lower than 1%, but the model-related instability is very large.

The same year, the same team looked for similar mode effects in a panel survey on victimization, where CAWI was introduces in addition to CATI. They found small and MAR selection (leading to unbiased estimates after post-stratification) that they could estimate using previous waves of the same panel. On the other hand, they found major mode measurement effects: double counts in violent events were more frequent on CAWI than on CATI. The same violent event was more often quoted twice by Internet: the survey described a series of violent situations, and some events or episodes had several dimensions, and were referred to more than once, despite the instruction to refer only once to each episode (Castell et al. 2023b). These double counts in violent events were more frequent on CAWI than on CATI: Internet respondents declared more often having been victim of violence or offenses, to questions on objective facts (theft, vandalism, physical violence) as well as on more subjective (insults, threats). An adaptation of the questionnaire so self-administration led to lower these mode effects.

These two recent examples show that, when response rate are low, selection effects can be large; in some cases mode measurement effects can also occur, due to satisficing or due to the overall ergonomic of the survey.

3 Estimating mode effects in GGS Round II surveys

3.1 Countries with more than one data collection mode

The following analyses are using two recently conducted GGS, the French test and the Uruguayan survey for wave 1 of GGS – Round II. Both surveys were multimode, and conducted on a representative sample of men and women aged 18-79.

In France, 831 questionnaires were collected in the 2021 test: 367 were conducted by telephone (CATI), and 431 on Internet (CAWI). 31 questionnaires were mixed, with 30 started in CAWI and completed later on CATI; this latter category is omitted in the analyses, which focus on comparing mode measurement effects in CAWI and CATI). The test was organized to test the questionnaire, compare CAWI and CATI modes, as well as the use of incentives (Bouchet-Valat et al. 2022). Our sample is thus made of 798 respondents, with 54% CAWI.

In Uruguay, a sample of homes were selected in urban areas (cities with population larger than 5 000 inhabitants. A respondent was selected, with 90% answering a face-to-face interview (CAPI) and 10% (randomly) assigned to CAWI (Pardo et al. 2022). The final sample if 7 245 respondents, with 14% CAWI.

In both countries results from the subsample who filled the questionnaire on CAWI can thus be compared to a similar sample who answered to an interviewer, on telephone in the French test (CATI), face-to-face in Uruguay (CAPI).

3.2 Mode selection effects in France and Uruguay

The structure of the samples differ with younger respondents on CAWI in both countries, but the differences are not very large. These mode selection effects are thus likely to be controlled with post-stratification. The GGP Methods Group will offer a standard post-stratification on a series of standard information, common to all countries (Jablonski et al. 2022). The construction of standard weights by the GGP Methods group, using design weights and producing post-stratification weights, based on margins for sex, age, region of residence, level of education, and marital status, will allow a precise analysis of mode selection effects.
In countries with multi-mode, it is possible to post-stratify each subsample, or to consider the propensity to answer through one mode or the other in the post-stratification (Rosenbaum and Rubin, 1983). Many other variables could eventually be used for post-stratification, like e.g. working status, number of children, and the weighting process is not completed yet for these samples.

Selection effects can also be controlled and measured by comparing survey results to external information. A good example of such a robustness check is the comparison of cohort fertility indicators for Estonia, Norway, Finland, Denmark, and Sweden (Leocádio et al. 2023). The results were very similar, proving the high quality of the data. If some differences had been found, they could have come from selection effects or measurement effects: for instance lower fertility estimates may come from selection of adults with no or few children, or from the omission of some children in the fertility questions.

3.3 Mode measurement effects in France and Uruguay

We will examine here a series of data quality indices, using the French 2021 test and the 2022 Uruguay survey, both having used CAWI as well as interviewers on independent samples, with telephone in France and face-to-face in Uruguay.
We measured several data quality indices and, for each, we ran a linear regression using similar control variables in both countries (reference categories in parentheses). In France: sex (woman), age (30-44), standard of living (medium categories), education (tertiary), household type (couple no kid), urban unit size (rural). In Uruguay, similar controls were used: sex, age, income, education, city size (same reference categories). For France, an additional control on incentives was introduced (ref. = no incentive), but this control did not show significant impact on measurement effects. Our variable of interest is the data collection mode: CAWI is used as the reference group; in France we compare CATI data collection mode; in Uruguay the compared data collection mode is CAPI.

Non-response to quantitative questions

The first index of data quality relates to ***partial non-response***. We chose the following set of questions, with quantitative answers on frequencies, satisfaction scales, numbers: dem09, dem12, dem37, dem42, lhi02, hhd01b, hhd12, gen15a, gen15b, gen16, gen30, gen39a, gen39b, gen40, gen47, gen54, fer16b, fer16c, fer21, fer23, fer24, wel01, wel06, wel07, wel08, wrk01, wrk07, wrk35, att02, att09, att10, att11b, att11d. The questionnaire and the precise text of the questions and answers related to each variable can be found in the standard GGS-Round II questionnaire, (Gauthier et al. 2021).
The version 3.1.1 is available at https://www.ggp-i.org/data/methodology/).

We counted for each respondent the number of questions that were applicable, and the number of questions with non-response (don't know, refusal). Our dependent variable is the ratio (percent) of questions with a non-response. The overall proportion of non-response is around 9% for these questions.
In both countries, non-response is more frequent through CAWI, while the interaction with an interviewer makes it easier to get an answer; the impact is larger in France, and significant in both countries (Table 2).

Table 2. Non-response to a set of quantitative questions (%)

	France, CATI	Uruguay, CAPI
CAWI	0	0
CATI or CAPI	***-9,6***	***-3,0***

Note: the lower the proportion of non-response, the better.
Differences significant at the 0.05 level are in ***italics bold***.

Non-response on quantitative questions

Our second index also relates to ***partial non-response***. We then chose another set of questions, with qualitative answers: dem26, dem38a, dem38b, dem38c, dem38d, dem38e, dem38f, dem38g, dem39a, dem39b, dem39c, dem39d, fer25a, fer25b, fer25c, fer25d, fer25e, fer25f, fer25g, fer25h, fer26a, fer26b, fer26c, fer26d, fer26e, fer26f, gen49, gen51, wrk20, wrk46, inc01, inc06, fer05.
We also counted for each respondent the number of questions that were applicable, and the number of questions with non-response (don't know, refusal). The proportion of non-response is lower than for quantitative questions, around 6%.

The analysis partly confirmed that in both countries non-response is more frequent through CAWI, the differences being smaller in Uruguay and not significant at the 5% level (Table 3).

Table 3. Non-response to a set of qualitative questions (%)

	France, CATI	Uruguay, CAPI
	France	Uruguay
CAWI	0	0
CATI or CAPI	***-7,6***	-0,9

Note: the lower the proportion of non-response, the better.
Differences significant at the 0.05 level are in ***italics bold***.

Proportion of "rare" answers

Our third data quality index relates to satisficing, and more specifically to one dimension, the probability to give a frequent answer. For a set of variables in the questionnaire, we first defined "rare" answers as categories with low frequencies. Our definition is based on two criteria: first, the answer was given by less than 10% respondents (one exception with 15%); second, the categories being sorted by decreasing frequency, the category is the less frequent, or the difference with the next category is much smaller than with the previous one (choice called the "elbow method"). We chose the same set of questions, with qualitative answers, and considered only questions with a response (for each individual, questions with a non-response were excluded): dem26, dem38a, dem38b, dem38c, dem38d, dem38e, dem38f, dem38g, dem39a, dem39b, dem39c, dem39d, fer25a, fer25b, fer25c, fer25d, fer25e, fer25f, fer25g, fer25h, fer26a, fer26b, fer26c, fer26d, fer26e, fer26f, gen49, gen51, wrk20, wrk46, inc01, inc06, fer05. The categories defined as "rare" are presented in Annex 7.1. The overall proportion of "rare" answers is 9% in both countries.

The analysis confirmed that the answers are more diversified through CAWI mode (Table 4): the proportion of rare answers is lower with an interviewer, meaning that the respondents are more likely to give a "normative" answer, more frequent in the final distribution of answers. Most of the "rare" answer may be considered as "less desirable", like DEM38 on couple disagreement (Schumann and Lück 2023) or FER25 on expected (negative) changes related to the birth of a child; it may also be the case for fertility impairments or extreme income values. This greater diversity of answers on CAWI may then be interpreted as a lower desirability bias, as shown by Schumann and Lück on questions related to relationship quality: couple satisfaction, conflict frequency and conflict style in the German GGS pilot. The CAWI answers are then considered to be of better quality, while CATI or CAPI answers given to an interviewer would suffer from a more severe desirability bias. Another interpretation would be that respondents on internet are more likely to express an overall dissatisfaction when they are facing their screen, be it related to a selection effect (dissatisfied individuals being more likely to answer on Internet) or to a measurement effect (related to the interaction with the computer or smartphone). Schork et al. (2021) found that, in the Luxembourgish Labour Force Survey, objective variables are not affected by mode-specific measurement bias, and that web participants report lower satisfaction-levels on subjective variables (Wage Adequacy and Job Satisfaction) than telephone participants. As satisfaction is likely to be considered as desirable by some respondents, the two effects are not easy to disentangle.

Table 4. Rare answer to a set of qualitative questions (%)

	France, CATI	Uruguay, CAPI
	France	Uruguay
CAWI	0	0
CATI or CAPI	***-1,9***	***-1,1***

Note: The more frequent the rare answers, the better.
Differences significant at the 0.05 level are in ***italics bold***.

Proportion of systematic answers

Our fourth data quality index also relates to satisficing, more specifically to the probability to give the same answer to all questions within a battery. For instance, DEM38 on disagreement with the partner is made of seven questions related to disagreements. Our set of batteries of questions is the following:

- ATT03 (attitudes on marriage and children): att03a, att03b, att03d, att03e, att03g, att03h, att03i, att03j
- ATT07 (opinion on gender roles): att07a, att07b, att07c, att07d, att07g
- INC04 (afford to purchase various items): inc04a, inc04b, inc04c, inc04d, inc04e, inc04f, inc04g, inc04h, inc04i, inc04j, inc04k
- WEL09 (loneliness index): wel09a, wel09b, wel09c, wel09d, wel09e, wel09f
- FER25 (changes related to the eventual birth of a child): fer25a, fer25b, fer25c, fer25d, fer25e, fer25f, fer25g, fer25h
- FER26 (conditions for having a child): fer26a, fer26b, fer26e, fer26f, fer26h
- HHD11 (sharing household tasks): hhd11a, hhd11b, hhd11c, hhd11d, hhd11e, hhd11f
- DEM38 (disagreement with partner): dem38a, dem38b, dem38c, dem38d, dem38e, dem38f, dem38g
- DEM39 (conflict style): dem39a, dem39b, dem39c, dem39d
- WEL11 (depression index): wel11a, wel11b, wel11c, wel11d, wel11e

For each battery, the measurement index is based on the variance of the answers given to the questions related to that set.
For some batteries, like DEM38, all questions go in the same direction, so that an overall disagreement lead respondents to give the same answer e.g. if the respondents has never disagreed with her/his partner on any subject); the same is true for INC04, FER25, HHD11, WEL11. In that case, variances are supposed to be small, in most cases, even in the absence of straight-lining.
On the contrary, other sets of questions (like WEL09 on loneliness) were constructed so that a general feeling of loneliness lead to different answers (e.g. "I can rely on people" and "I miss having people around"). Similarly, ATT07 values on gender roles questions offer different answers for respondents attached to differentiated gender roles ("men" or "women"), as well as DEM39 on conflict style for systematic opinions. For these sets, a small variance is more likely to come from straight-lining, as giving the same answer to all questions appears less consistent in terms of opinions.

Table 5. Variance in the answers to a set of questions

	France, CATI	Uruguay, CAPI
	France	Uruguay
Reference for all regressions		
CAWI	0	0
a. ATT03: attitudes on marriage and children		
CATI or CAPI	-0,08	***-0,22***
b. ATT07: opinion on gender roles		
CATI or CAPI	0,00	***0,06***
c. INC04: afford to purchase various items		
CATI or CAPI	-0,01	-0,69
d. WEL09: loneliness index		
CATI or CAPI	***0,12***	***0,12***
e. FER25: changes related to the eventual birth of a child		
CATI or CAPI	0,08	0,00
f. FER26: conditions for having a child		
CATI or CAPI	***0,41***	
g. HHD11: sharing household tasks		
CATI or CAPI	0,15	
h. DEM38: disagreement with partner		
CATI or CAPI	-0,02	-0,01
i. DEM39: conflict style		
CATI or CAPI	***0,25***	***0,21***
j. WEL11 (depression index)		
CATI or CAPI	0,04	0,02

Note: the larger the variance of answers, the better.
Differences significant at the 0.05 level are in ***italics bold***.

The analysis did not confirm a systematic measurement effect related to straight-lining (Table 5). Nevertheless, the answers are more diversified (larger variance) in both countries for two sets with no systematic answers: loneliness index (Table 5.d) and conflict style (Table 5.i). In France a measurement effect appears significant for a third set (Table 5.f, conditions for having a child, not asked in Uruguay); in Uruguay, two additional sets present measurement effects (Table 5.a and .b, attitudes on marriage and children, and opinions on gender role; but the former with less diversity on CAPI, and thus more straight-lining with an interviewer, which is counter-intuitive and could be related to specific desirability bias on these questions in Uruguay.

3.4 Data quality on CAWI is not lower than on CAPI or CATI

From the comparisons presented above, the conclusion regarding data quality is not straightforward. GGS data from France (CAWI and CATI) and Uruguay (CAWI and CAPI) confirm than non-response is more frequent on CAWI. This is related to the possibility to go to the next questions (and the next screen) without answering a question. Tests have

shown that respondents feel offended and upset if they are forced to answer a question to pursue. As no interviewer is present, the risk of breakoff is then large. Tests have shown that breakoffs are frequent on CAWI. The GGS questionnaire is long, and breakoffs were frequent in the GGP Pilot: 17%, more than 20% in Germany and Portugal (Emery et al. 2022). Breakoffs are spread over the questions. Incentives limit the proportion of incomplete questionnaires. In the French GGS, the interviewers will remind respondents who started to fill in the questionnaire on CAWI but did not go to the end, and will offer them to pursue on telephone or on internet.

Another issue related to the data collection process is related to finding the good respondent: in the French test, the comparison between the original sample (from tax data) and the respondents showed that 1,9% of CAWI respondents were likely not the sampled individual, but another member of their household contacted by telephone or mail. Such error did not occur on CATI. Choosing the respondent from a sample of households is complicated on CAWI, and can lead to manty errors.

Regarding straight-lining, the mode effect is not systematic (and not larger for the batteries at the end of the questionnaire), but in all the risk of systematic answers (and thus low quality indices) is larger on CAWI.

Considering the diversity of answers, "rare" answers are more frequent on CAWI, which can be interpreted as a smaller desirability bias: CAWI being a self-administered mode, less socially desirable reflections would be more easily expressed (Schumann and Lück 2023). Further comparisons could be made for countries where PAPI mode was used, in order to check whether the same mode measurement effect can be observed. In Germany, PAPI was chosen because the COVID-19 epidemics made face-to-face interviews impossible. In Sweden it has been considered a cheap data collection mode, complementary to CAWI. According to Klausch et al. (2013), the paper-and-pencil survey (PAPI mode, based on questionnaires sent by postal mail) offers results consistent with CAWI, as both modes are self-administered, and is complementary with CAWI in germs of potential respondents. As CAWI and PAPI have been combined in Sweden and Germany, additional mode measurement effects could be tested from these country GGS datasets.

All in all, CAWI appears to be a data collection mode leading to high quality data, as far as non-response are not too frequent and straight-lining is avoided. Further progress in questionnaire ergonomics on laptops and smartphone may increase data quality. The long duration of the GGS questionnaire does not appear as a major problem for CAWI data collection mode (Emery et al. 2022).

The assessment of data quality can be internal or external (Leocádio et al. 2023). Furthermore, data quality assessment can be based on univariate distributions or on more complex relations. For instance, Piccitto et al. (2022), using GGS data from Croatia and Germany, show that "Although respondents report lower subjective well-being in web than in face-to-face mode, the relationships between these variables and a range of objective and subjective indicators are relatively stable". Thus, a mode measurement effect on the univariate distribution is present (which could even considered as the consequence of lower desirability bias on CAWI), but the main result is that the structural important results are less subject to mode effects than the univariate distributions.

4 Practical guidelines

The GGS – Round II data collection is more and more based on push-to-web, using CAWI questionnaires. Other data collection mode are used as complementary in some countries: CAPI, CATI and PAPI questionnaires were or will be used in some countries. For countries using a multi-mode data collection, the following guidelines can be useful.

4.1 Check for mode selection effects

In case several data collection modes are used, mode selection effects can be checked with precisely analysing inclusion probability for each mode, starting with basic demographic variables used for the standard post-stratification.

4.2 Post-stratification of the sample

The starting point is the post-stratification proposed by the GGP Methods group, based on the margins for a few variables available in all countries: sex, age, region, level of education, marital status. Of course, this requires using the same definitions for the GGP and the population margins. This is a very efficient first step, partly controlling for selection effects (irrespective of mode). It can be completed in many ways, when additional information is available:

- Using crossed-margins, like e.g. education by age, marital status by sex, etc.
- Eventually use other margins (other variables, like e.g. professional status, country of birth)
- and eventually some variables under study, like parity (number of children ever born), if external data are available

These additional checks could lead to additional weights. In any case, all weights should conform to the Methods group recommendations, so that the margins on the five proposed variables are respected.

4.3 Check for mode response effects

Mode measurement effects can be checked, using the program included in annex 7.2.

As a robustness check, the information on data collection mode could systematically be included in the multivariate analyses, in order to check for the absence of main effects and interactions.

5 Next steps

As a conclusion, some thoughts and questions for the future are presented. Results of further analyses from the national teams should be shared, and a methods seminar could be organized within the GGP infrastructure.

For now, most national teams are collecting GGS data with one only data collection mode. For international comparisons, controlling for data collection mode may be complicated and not necessarily useful. A standard post-stratification will hopefully control for most selection effects, and measurement effects may come from data collection mode, but also in some cases to translation, adaptation, data collection process in general, social context at the time of survey, etc. Furthermore, mode effects are likely to be more limited in multivariate models, but this has to be confirmed by further analyses.

Sequential multimode is likely to develop (CAWI followed by PAPI, CATI ou CAPI), in order to increase response rates and maximize the representativeness of the sample. As the response rates are low, mode effects may not be the main issue at stake. Other selection bias (differential response rates) may be present for all modes, leading to post-stratification of the whole sample in all countries, irrespective of the mode of data collection. The question of the reference mode may then vanish, as each subsample may not be representative, but the whole sample, putting together all questionnaires, irrespective of the mode of data collection.

Estimating accurately mode effects will be possible in a very limited number of countries, where mode sub-samples were defined a priori. In most countries with one mode only (or with competing data collection mode (the respondents choosing to answer through one or the other mode), or sequential modes (non-respondents on CAWI being contacted by postal mail or by telephone), mode effects may be difficult to estimate.

As mode effects can be large for some variables, and almost null for others, an overall correction of mode effects may lead to increase the variance of weights without much gain. Regarding selection into the sample, specific efforts to include specific subgroups like migrants or lower educated persons, should be encouraged. When a sample frame is available, the sample can be extracted with unequal probabilities, in order to compensate differential response rate sand limit the variance of the weights used in the post-stratification.

Another possibility is to produce specific datasets adjusted to control for mode effects, thus allowing for more accurate international comparisons, especially for variables for which mode effects are suspected (like e.g. disagreements within the couple). Other methods could also be considered. In order to be compliant to GDPR regulations, we could eventually change the value of some variable for some respondents. This could be done through imputation changing some “unlikely” answers for some individuals, and replacing them by a more likely answer in the preferred mode was used. These “model based imputations” would introduce some changes in the data, thus making them “more anonymous”. This could also be done in order to partly “correct” for mode effects in some cases. Of course we have to be careful on that, but imputations can be useful for correcting for non-response, adjusting for mode effects, introducing some noise in the data (in the sense that identification will be more difficult or less harmful). Preliminary trials have been conducted (Legleye et al. 2018; Barret and Cissé 2018), but the method still needs to be assessed.

6 Bibliography

Barret C ., Cissé M (2018), Agrégation de données multimode : impact sur la modélisation des variables présentant un effet de mesure [Multimode data aggregation: impact on the modeling of variables with a measurement effect], Insee : journées de méthodologie statistique, session 22, http://www.jms-insee.fr/2018/S22_3_ACTEv2_BARRET_JMS2018.pdf

Biemer P. (2001), Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing, *Journal of Official Statistics*, 17, p. 295-320

Bouchet-Valat M., Bondon M., Breda-Popa R., Cochet P., Charrance G., Markou E., Toulemon L. (2022), Telephone, Internet, and Incentives: First Lessons from the French GGP2020 Pilot. GGP-Connect Seminar n°8, https://www.ggp-i.org/ggp-connect-seminar-series/#toc13. Presented as a poster at the EPC 2022 General Conference, Poster 1.14. See long abstract at https://epc2022.eaps.nl/uploads/210863

Castell L., Favre-Martinoz C., Paliod N., Sillard P. (2023a), Redressements de la première vague de l'enquête EpiCov : un exemple de correction des effets de sélection dans les enquêtes multimodes [Adjustments to the first wave of the EpiCov survey: an example of correcting for selection effects in multi-mode surveys]. Insee, *Documents de travail*, n° M2023/02, https://www.insee.fr/fr/statistiques/7452990

Castell L., Clerc M., Croze D., Legleye S., Nougaret A. (2023b), Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête Cadre de Vie et Sécurité [Declared victimizations and mode effects: lessons from the experimental multi-mode panel of the Living Environment and Safety survey]. Insee, *Documents de travail*, n° M2023/02, https://www.insee.fr/fr/statistiques/7709248

Emery T., Cabaco S., Fadel L., Lugtig L., Toepoel V., Schumann A., Lück D., Bujard M. (2022), Breakoffs in an hour long online survey. SocArXiv. DOI: 10.31235/osf.io/ja8k4

Gauthier, A. H., Liefbroer, A., Ajzen, I., Aassve, A., Beets, G., Billari, F., Bühler, C., Bujard, M., Cabaço, S., Corijn, M., Désesquelles, A., Dommermuth, L., Dykstra, P., Emery, T., Fadel, L., Fokkema, T., Hansen, T., Hlebec, V., Hoem, J., Klobas, J., Kogovšek, T., Koops, J. C., Kveder, A., Lappegård, T., Lück, D., Lugtig, P., MacDonald, A., Macura, M., Makay, Z., Mills, M. C., Murinkó, L., Mynarska, M., Neyer, G., Pailhé, A., Petrič, G., Pinnelli, A., Ratikainen, J., Rayboud, A., Rijken, A., Slagsvold, B., Solaz, A., Spéder, Z., Thévenon, O., Vikat, A. (2021). Generations and Gender Survey Baseline Questionnaire 3.1.1 Netherlands Interdisciplinary Demographic Institute. https://www.ggp-i.org/data/methodology/

Groves, R., Singer E. Corning A. (2000), Leverage-Saliency Theory of Survey Participation: Description and an Illustration, *Public Opinion Quarterly* 64(3): 299-308.

Groves R., Fowler F. Jr., Couper M., Lepkowski J., Singer E., Tourangeau R. (2004), *Survey Methodology* (Wiley Series in Survey Methodology).

Heckman J. (1979), Sample Selection Bias as a Specification Error, *Econometrica*, 47(1): 153-161.

Hox J., De Leeuw E., Klausch T. (2015), Mixed Mode Research: Issues in Design and Analysis *in* Biemer et al., *Total Survey Error in Practice*, John Wiley & Sons, pp.511-530.

Jablonski W., Liefbroer A., Leocádio V.A. & the GGS Methods group (2023), Data quality in the new GGS-II, GGP-Connect seminar n°14, https://www.ggp-i.org/ggp-connect-seminar-series/#toc7

Klausch T. (2014), *Informed Design of Mixed-Mode Surveys: Evaluating mode effects on measurement and selection error*, PhD Thesis, Department of Methodology and Statistics. Ultrecht University Repository. https://dspace.library.uu.nl/handle/1874/300673

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

Krosnick J. (1991), Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5: 213-236.

Legleye S., de Perretti G., Razafindranovona T. (2018) Agréger les échantillons d'une enquête multimode en limitant l'effet de mesure: une proposition d'imputation raisonnable et pragmatique [Aggregating multimode survey samples while limiting measurement effect: a reasonable and pragmatic imputation proposal]. Insee : journées de méthodologie statistique, session 22, http://www.jms-insee.fr/2018/S22_2_ACTE_LEGLEYE_JMS2018.pdf

Leocádio V., Gauthier A., Mynarska M., Costa R. (2023), The quality of fertility data in the web-based Generations and Gender Survey; *Demographic Research* 49(3): 31–46. See also the GGP-Connect seminar n°14, https://www.ggp-i.org/ggp-connect-seminar-series/#toc7

Lugtig P., Toepoel V., Emery T., Cabaço S., Bujard M., Naderi R., Schumann A. Lück D (2022) Can we Successfully Move a Cross-national Survey online? Results from a Large Three-country Experiment in the Gender and Generations Programme survey. Available at https://osf.io/preprints/socarxiv/mu8jy.

Pardo I., Cabella W., Fernández Soto M., Pedetti G., Klaczko I., Pelufo S., Anzalone F., Sala M. (2022), Conducting the GGS in Uruguay. Fieldwork challenges. GGP-Connect seminar, n°10, https://www.ggp-i.org/ggp-connect-seminar-series/#toc11

Piccitto G., Liefbroer A., Emery T. (2022), Does the survey mode affect the association between subjective well-being and its determinants? An experimental comparison between face-to-face and web mode. *Journal of Happiness Studies* 23 (7): 3441-3461

Rosenbaum P., Rubin D. (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika* 70(1): 41-55.

Schork, J. Riillo C., Neumayr J. (2021), Survey Mode Effects on Objective and Subjective Questions: Evidence from the Labour Force Survey, *Journal of Official Statistics* 37(1): 213-237.

Schouten B., Bethlehem J., Beullens K., Kleven Ø., Loosveldt G., Luiten A., Rutar K., Shlomo N., Skinner C. (2012), Evaluating, Comparing, Monitoring, and Improving

Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review* 80(3): 382-399.

Schumann A., Lück D. (2023), Better to ask online when it concerns intimate relationships? Survey mode differences in the assessment of relationship quality. *Demographic Research* 48(22): 609-640

Warszawski J., Beaumont A.-L., Seng R., Lamballerie X. de, Rahib D., Lydié N., Slama R., Durrleman S., Raynaud P., Sillard P., Beck F., Meyer L., Bajos N. & The EPICOV study group (2022), Prevalence of SARS-Cov-2 antibodies and living conditions: the French national random population-based EPICOV cohort, *BMC Infectious Diseases* 22: 41.

7 Annex

7.1 Annex 1. Definition of rare modalities

The "rare" modalities are shaded in light blue

FRANCE

dem26	n	val%
2	359	58.0
1501	120	19.4
3	61	9.9
5	28	4.5
8	16	2.6
1	14	2.3
12	8	1.3
11	4	0.6
7	3	0.5
9	3	0.5
10	2	0.3
4	1	0.2

dem38a	n	val%
2	220	35.8
3	171	27.9
1	143	23.3
4	64	10.4
5	16	2.6

dem38b	n	val%
1	278	45.1
2	187	30.4
3	112	18.2
4	26	4.2
5	13	2.1

dem38c	n	val%
1	209	33.9
2	201	32.6
3	153	24.8
4	44	7.1
5	9	1.5

URUGUAY

dem26	n	val%
2	2065	50,3
6	768	18,7
3	594	14,5
5	232	5,6
8	167	4,1
1	126	3,1
12	73	1,8
7	34	0,8
4	24	0,6
11	24	0,6
9	1	0,0

dem38a	n	val%
1	1535	38,1
2	1081	26,8
3	948	23,5
4	340	8,4
5	130	3,2

dem38b	n	val%
1	2171	53,7
2	891	22,0
3	700	17,3
4	211	5,2
5	72	1,8

dem38c	n	val%
1	2294	56,5
2	843	20,8
3	694	17,1
4	187	4,6
5	41	1,0

dem38d	n	val%
1	290	47.1
2	206	33.4
3	96	15.6
4	15	2.4
5	9	1.5

dem38d	n	val%
1	2973	73,0
2	666	16,4
3	345	8,5
4	67	1,6
5	19	0,5

dem38e	n	val%
1	260	45.4
2	137	23.9
3	128	22.3
4	37	6.5
5	11	1.9

dem38e	n	val%
1		72,4
2	521	14,4
3	331	9,1
4	111	3,1
5	35	1,0

dem38f	n	val%
1	420	78.4
2	62	11.6
3	39	7.3
4	13	2.4
5	2	0.4

dem38f	n	val%
1	2922	84,4
2	277	8,0
3	170	4,9
4	53	1,5
5	39	1,1

dem38g	n	val%
1	229	42.3
3	132	24.4
2	128	23.7
4	42	7.8
5	10	1.8

dem38g	n	val%
1	1912	58,1
3	583	17,7
2	580	17,6
4	164	5,0
5	52	1,6

dem39a	n	val%
3	208	34.0
2	193	31.6
1	154	25.2
4	47	7.7
5	9	1.5

dem39a	n	val%
3	1144	28,4
1	951	23,6
2	949	23,6
4	712	17,7
5	273	6,8

dem39b	n	val%
4	234	38.4
3	154	25.2
5	134	22.0
2	68	11.1

dem39b	n	val%
4	1417	35,2
3	1058	26,3
5	701	17,4
2	509	12,6

1	20	3.3

1	343	8,5

dem39c	n	val%
2	191	31.1
3	189	30.7
1	166	27.0
4	59	9.6
5	10	1.6

dem39c	n	val%
1	2164	53,3
2	1037	25,5
3	648	16,0
4	167	4,1
5	44	1,1

dem39d	n	val%
1	248	40.4
2	203	33.1
3	128	20.8
4	31	5.0
5	4	0.7

dem39d	n	val%
1	2274	56,1
2	804	19,8
3	691	17,1
4	203	5,0
5	80	2,0

fer25a	n	val%
5	171	32.1
4	168	31.5
3	153	28.7
2	28	5.3
1	13	2.4

fer25a	n	val%
3	1572	43,2
4	1204	33,1
5	393	10,8
2	343	9,4
1	126	3,5

fer25b	n	val%
4	223	41.7
3	147	27.5
5	139	26.0
2	20	3.7
1	6	1.1

fer25b	n	val%
4	1684	46,1
3	1183	32,4
5	531	14,5
2	185	5,1
1	67	1,8

fer25c	n	val%
3	184	34.8
4	171	32.3
5	84	15.9
2	65	12.3
1	25	4.7

fer25c	n	val%
3	1684	46,1
4	1037	28,4
2	501	13,7
5	266	7,3
1	165	4,5

fer25d	n	val%
2	198	37.6
3	123	23.4

fer25d	n	val%
3	1480	40,2
2	1103	30,0

1	96	18.3
4	66	12.5
5	43	8.2

4	523	14,2
1	418	11,4
5	154	4,2

fer25e	n	val%
3	224	45.8
4	154	31.5
5	81	16.6
2	23	4.7
1	7	1.4

fer25e	n	val%
3	1908	53,9
4	1010	28,5
2	264	7,5
5	256	7,2
1	102	2,9

fer25f	n	val%
3	232	52.0
4	124	27.8
5	58	13.0
2	23	5.2
1	9	2.0

fer25f	n	val%
3	1748	63,2
4	591	21,4
2	209	7,6
5	132	4,8
1	86	3,1

fer25g	n	val%
3	250	49.2
2	155	30.5
1	43	8.5
4	31	6.1
5	29	5.7

fer25g	n	val%
NO		

fer25h	n	val%
3	254	52.3
2	91	18.7
4	74	15.2
5	37	7.6
1	30	6.2

fer25h	n	val%
NO		

fer26a	n	val%
5	204	36.3
4	178	31.7
3	68	12.1
2	65	11.6
1	47	8.4

fer26a	n	val%
4	1415	36,5
5	797	20,6
3	699	18,0
2	484	12,5
1	479	12,4

fer26b	n	val%

fer26b	n	val%

5	232	42.0
4	144	26.0
3	69	12.5
2	60	10.8
1	48	8.7

4	1399	36,5
5	1037	27,1
3	590	15,4
2	416	10,9
1	386	10,1

fer26c	n	val%
4	197	35.2
5	189	33.8
3	82	14.6
1	52	9.3
2	40	7.1

fer26c	n	val%
NO		

fer26d	n	val%
1	206	36.9
2	101	18.1
5	101	18.1
4	90	16.1
3	61	10.9

fer26d	n	val%
NO		

fer26e	n	val%
5	194	36.7
1	112	21.2
4	108	20.4
3	71	13.4
2	44	8.3

fer26e	n	val%
5	1211	32,5
4	1103	29,6
3	720	19,3
1	394	10,6
2	296	7,9

fer26f	n	val%
4	170	32.9
5	127	24.6
3	88	17.1
1	74	14.3
2	57	11.0

fer26f	n	val%
4	1421	38,0
3	759	20,3
5	735	19,7
2	432	11,6
1	389	10,4

gen49	n	val%
1	164	25.9
1503	109	17.2
1502	69	10.9
7	58	9.2
5	49	7.8

gen49	n	val%
1	3045	50,9
2	866	14,5
3	807	13,5
0	485	8,1
6	293	4,9

1506	46	7.3
1508	38	6.0
1504	28	4.4
1501	24	3.8
0	23	3.6
8	17	2.7
1505	6	0.9
1507	1	0.2

4	231	3,9
5	202	3,4
7	35	0,6
8	24	0,4

gen51	n	val%
1	180	26.9
1503	108	16.2
1506	73	10.9
1502	66	9.9
1508	55	8.2
5	55	8.2
1501	35	5.2
7	33	4.9
0	26	3.9
1504	19	2.8
8	12	1.8
1505	5	0.7
1507	1	0.1

gen51	n	val%
1	3294	50,7
2	947	14,6
3	884	13,6
0	464	7,1
4	400	6,2
6	263	4,0
5	197	3,0
7	34	0,5
8	16	0,2

wrk20	n	val%
1	233	55.3
2	188	44.7

wrk20	n	val%
1	1866	67,5
2	900	32,5

wrk46	n	val%
1	190	51.2
2	181	48.8

wrk46	n	val%
1	1123	58,6
2	794	41,4

inc01	n	val%
83	188	25.9
9	103	14.2
10	95	13.1
8	79	10.9
11	74	10.2
7	69	9.5
6	43	5.9

inc01	n	val%
4005	984	31,5
4006	913	29,2
4004	546	17,5
4007	202	6,5
4001	171	5,5
4003	163	5,2
4002	97	3,1

12	22	3.0
5	22	3.0
1	13	1.8
4	9	1.2
2	5	0.7
3	3	0.4

4008	51	1,6

inc06	n	val%
6	162	23.6
5	128	18.7
4	125	18.2
3	83	12.1
7	74	10.8
2	40	5.8
8	27	3.9
1	24	3.5
9	22	3.2

inc06	n	val%
4001	1598	25,7
4002	1149	18,5
4008	916	14,7
4003	785	12,6
4004	671	10,8
4005	484	7,8
4006	368	5,9
4007	256	4,1

fer05	n	val%
3	176	44.0
4	171	42.8
1	31	7.8
2	22	5.5

fer05	n	val%
4	1602	47,5
3	1071	31,8
1	505	15,0
2	195	5,8

7.2 Annex 2. R code on mode measurement effects

This code will be completed for studying other mode measurement effects, updated and offered in the GGP-i website.

Structure variables (definition from our administrative file)

Scenario (data collection mode), Incentive, Sex, Age, Standard of living, Education, Household type, Urban unit size,

```
# data collection mode
    DF$modedet_r <- relevel(as.factor(DF$modedet_r), "CAWI")
# Incentive
    DF$incentive <- relevel(as.factor(DF$incentive), "None")
# Sex
    DF$sexe_co_r <- relevel(as.factor(DF$sexe_co_r), "Woman")
# Age
    DF$age_fideli_r <- relevel(as.factor(DF$age_fideli_r), "[30,45)")
# Standard of living
    DF$decile_nivviem_n_1_r <- relevel(as.factor(DF$decile_nivviem_n_1_r), "Deciles 5-7")
# Education
    DF$dem07_r <- relevel(as.factor(DF$dem07_r), "Etudes supérieures")
# Household type
    DF$type_menf_r <- relevel(as.factor(DF$type_menf_r), "Couple no kid")
# Urban unit size
    DF$TUU2017_r <- relevel(as.factor(DF$TUU2017_r), "Communes rurales")
```

Response effects indices

- Non-response

1. ***Quantitative variables used to count the number of non-responses***

```
# list of variables
"dem09", "dem12", "dem37", "dem42", "lhi02", "hhd01b", "hhd12", "gen15a", "gen15b",
"gen16", "gen30", "gen39a", "gen39b", "gen40", "gen47", "gen54", "fer16b", "fer16c",
"fer21", "fer23", "fer24", "wel01", "wel06", "wel07", "wel08", "wrk01", "wrk07", "wrk35",
"att02", "att09", "att10", "att11b", "att11d"
```

Code to model propensity for partial non-response

```
    # Check non-response a or b, refuse, don't know (by default NA, to be recoded)
    for (i in 1:length(quantis_nr)){ # array with all quantitative variables
      DF[,quantis_nr[i]] <- coalesce(na_tag(DF[[quantis_nr[i]]]),
    as.character(DF[[quantis_nr[i]]]))
    }
    # number of response items (!= NA)
    DF$d_quanti <- rowSums(!is.na(DF2[,quantis_nr]))
    # number of non-response items (!= NA) type a or b (don't know, refuse)
    DF$n_quanti <- rowSums(DF2[,quantis_nr]=="a" | DF2[,quantis_nr]=="b", na.rm=T)
    # proportion of non-response as an index of partial non-response
    DF$nr_quanti <- as.numeric(DF$n_quanti / DF$d_quanti)*100
    # regression on that index
```

```
reg <- lm(nr_quanti ~ sexe_co_r + age_fideli_r + decile_nivviem_n_1_r + dem07_r
+ TUU2017_r + type_menf_r + modedet_r + incentive , data=DF)
ggcoef_model(reg, point_size=3.2)
```

2. ***Qualitative variables used to count the number of non-responses***

 # list of variables
"dem26", "dem38a", "dem38b", "dem38c", "dem38d", "dem38e", "dem38f", "dem38g", "dem39a", "dem39b", "dem39c", "dem39d", "fer25a", "fer25b", "fer25c", "fer25d", "fer25e", "fer25f", "fer25g", "fer25h", "fer26a", "fer26b", "fer26c", "fer26d", "fer26e", "fer26f", "gen49", "gen51", "wrk20", "wrk46", "inc01", "inc06", "fer05"

Code to model propensity for partial non-response (similar to the code for quantitative variables)

```
# Check non-response a or b, refuse, don't know (by default NA, to be recoded)
for (i in 1:length(qualis)){ # array with all qualitative variables
   DF[,qualis[i]] <- coalesce(na_tag(DF[[qualis[i]]]), as.character(DF[[qualis[i]]]))
}
Code 83 for variable inc01, recoded as non-response type b
DF$inc01 <- fct_recode(DF$inc01,
 "b" = "83")
# number of response items (!= NA)
DF$d_quali <- rowSums(!is.na(DF2[,qualis]))
# number of non-response items (!= NA) type a or b (don't know, refuse)
DF$n_quali <- rowSums(DF2[,qualis]=="a" | DF2[,qualis]=="b", na.rm=T)
# proportion of non-response as an index of partial non-response
DF$nr_quali <- as.numeric(DF$n_quali / DF$d_quali)*100
# regression on that index
reg <- lm(nr_quali ~ sexe_co_r + age_fideli_r + decile_nivviem_n_1_r + dem07_r +
TUU2017_r + type_menf_r + modedet_r + incentive , data=DF)
ggcoef_model(reg, point_size=3.2)
```

- Rare answers

 # All qualitative variables are transformed into proxies with value 1 in case of a rare answer and 0 otherwise (base DF3)

Rare answers are defined as categories with few answers. The limit can be settled from the shape of the distribution. The "elbow method" involves identifying the points at which the frequency of answers (ordered in decreasing order) decreases much more slowly as the density function is traversed from left to right. The attached document rare_modalities.xlsx shows the definition of rare answers.

```
# number of qualitative variables with a positive response (!= NA)
DF3$d_quali <- rowSums(!is.na(DF3[,qualis]))
# number of qualitative variables with a rare answer
DF3$d_quali <- rowSums(DF3[,qualis]==1, na.rm=T)
# proportion of rare answers as an index of non-normative answers
DF3$rares <- as.numeric(DF3$n_quali / DF3$d_quali)*100
# regression on that index
reg <- lm(rares ~ sexe_co_r + age_fideli_r + decile_nivviem_n_1_r + dem07_r +
TUU2017_r + type_menf_r + modedet_r + incentive , data=DF3)
ggcoef_model(reg, point_size=3.2)
```

- Straight-lining

 # list of the variables contained in each set of questions

ATT03

"att03a","att03b","att03d","att03e","att03g","att03h","att03i","att03j"

ATT07

"att07a","att07b","att07c","att07d","att07g"

INC04

"inc04a","inc04b","inc04c","inc04d","inc04e","inc04f","inc04g","inc04h","inc04i","inc04j","inc04k"

WEL09

"wel09a", "wel09b", "wel09c", "wel09d", "wel09e", "wel09f"

FER25

'fer25a', 'fer25b', 'fer25c', 'fer25d', 'fer25e', 'fer25f', 'fer25g', 'fer25h'

HHD11

'hhd11a', 'hhd11b', 'hhd11c', 'hhd11d', 'hhd11e', 'hhd11f'

DEM38

'dem38a', 'dem38b', 'dem38c', 'dem38d', 'dem38e', 'dem38f', 'dem38g'

DEM39

'dem39a','dem39b','dem39c','dem39d'

WEL11

'wel11a','wel11b','wel11c','wel11d','wel11e'

Code to model propensity for straight-lining

 # The first step is to estimate the variance of the answers given among a set of questions (INC04 in this example)

```
temp <-DF %>%
  select(inc04) %>% # array with all variables in the set INC04
  t()
# variance of answers added in the original file (DF)
vec <- as.data.frame(apply(temp, 2, var, na.rm=T))
DF$vec_inc04 <- as.numeric(vec$`apply(temp, 2, var, na.rm = T)`)
# regression on that index
reg <- lm(vec_inc04 ~ sexe_co_r + age_fideli_r + decile_nivviem_n_1_r + dem07_r + TUU2017_r + type_menf_r + modedet_r + incentive , data=DF)
ggcoef_model(reg, point_size=3.2)
```

 # The same method is used for each set of questions