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

DEMOGRAPHIC RESEARCH

VOLUME 48, ARTICLE 22, PAGES 609–640
PUBLISHED 4 MAY 2023
https://www.demographic-research.org/Volumes/Vol48/22/
DOI: 10.4054/DemRes.2023.48.22

Research Article

Better to ask online when it concerns intimate relationships? Survey mode differences in the assessment of relationship quality

Almut Schumann

Detlev Lück

Contents

Demographic Research: Volume 48, Article 22
Research Article

Better to ask online when it concerns intimate relationships? Survey mode differences in the assessment of relationship quality

Almut Schumann[1]

Detlev Lück[2]

Abstract

BACKGROUND
The assessment of relationship quality is a key construct in family research and relies on several indicators. As answer behavior for sensitive and subjective questions can be biased by the interview situation, the emerging switch from face-to-face mode to web or mixed mode in surveys challenges the comparability of measurements.

OBJECTIVE
This study investigates the impact of two modes of data collection – face-to-face mode and web mode – on central measurements of relationship quality in quantitative family research.

METHODS
In a German experimental pilot study (2018) within the Generations and Gender Programme, target persons were randomly assigned to face-to-face or online interviews. Mode differences are assessed by comparing distributions for various indicators of relationship quality. To adjust for confounders, post-stratification weighting and multivariate regression analysis are applied.

RESULTS
Findings reveal consistent mode effects for almost all indicators of relationship quality even after adjusting for confounders. Respondents in web mode assess their relationship quality substantially lower than respondents in face-to-face mode, thinking more often about breaking up and reporting lower satisfaction and more conflicts.

CONCLUSION
Web mode seems to support less socially desirable reflections on respondents' relationships compared to face-to-face mode. Family researchers should consider survey

[1] Federal Institute for Population Research (BiB), Wiesbaden, Germany.
Email: Almut.Schumann@bib.bund.de.
[2] Federal Institute for Population Research (BiB), Wiesbaden, Germany.

design decisions when evaluating intimate relationships, particularly in longitudinal and cross-national studies.

CONTRIBUTION

Findings on the assessment of relationships in family research based on self-administered modes, such as web mode, can be considered more reliable than those based on interviewer-administered modes.

1. Introduction

The assessment of relationship quality is one of the most frequently addressed topics in research on intimate relationships (see Bradbury, Fincham, and Beach 2000; Fincham and Beach 2006; Karney and Bradbury 2020). Relationship quality is connected with the stability of relationships and therefore often serves as a predictor for processes such as separation, family formation, and marriage (Karney and Bradbury 1995; Lewis and Spanier 1979). Not only family sociologists and demographers but also family psychologists frequently use indicators for relationship quality as determinants for various outcomes. Central and frequently analyzed indicators for relationship quality in quantitative family research are subjectively perceived stability (e.g., van Damme and Dykstra 2018; Wiik, Keizer, and Lappegard 2012), satisfaction with the relationship (e.g., Arránz-Becker 2013; Schmid et al. 2021), and certain interactions between partners, such as conflict behavior (e.g., Huß and Pollmann-Schult 2020; Kluwer and Johnson 2007).

Most of the studies mentioned compare aspects of relationship quality between different countries and cultural backgrounds or between different points in time or life course phases (e.g., Wiik, Keizer, and Lappegard 2012; Huß and Pollmann-Schult 2020; Schmid et al. 2021). Whenever different data sources are used for longitudinal or cross-national studies, analyses strongly depend on consistently high data quality and the comparability of data. Limitations on reliability and comparability can have many causes. A particularly important determinant is the mode of data collection (Groves et al. 2004). Face-to-face interviews have for decades been the most common mode of data collection for large-scale survey programs in the landscape of social science and family research, mainly because of their comparably high response rate and good coverage for the achievement of population-representative samples (De Leeuw, Hox, and Dillman 2008; Groves et al. 2004). Notwithstanding, one of the known and well-researched downsides of face-to-face interviews is that the personal interview situation supports the underreporting of sensitive topics, such as illicit or sexual behavior (Aquilino 1991; Tourangeau and Smith 1996).

However, this conventional wisdom is currently undergoing a reassessment. For data collectors, the switch from traditional personal interviews to online interviews is becoming increasingly attractive because web interviews are much more cost-efficient than face-to-face interviews, especially in countries where labor costs for interviewers are high (Bethlehem and Biffignandi 2012). Additionally, they facilitate rapid data collection and delivery (Couper 2011). Moreover, the ubiquity of mobile phones and smartphones is leading to an increasing use of mobile devices to complete web surveys (Gummer et al. 2023), giving respondents easier access to surveys – via QR codes, for example – and allowing them to answer at any time or from any place (Couper, Antoun, and Mavletova 2017). Last but not least, the COVID-19 pandemic strongly accelerated this transition by forcing established face-to-face studies, such as the Generations and Gender Survey (GGS) and the German family panel pairfam, to switch to web interviews (Gummer et al. 2020). So we may currently be witnessing the establishment of the self-administered online interview as a new standard mode of data collection, at least in Western Europe and in other countries with high labor costs and appropriate sampling frames.

In view of this development, the potential impact of the mode of data collection on data quality and on substantive analyses has become an even more relevant question for empirical analysis. We examine two survey modes that mark the starting and ending points of the transition described above: face-to-face interviews and web interviews. These two modes show the greatest difference in interviewer involvement: web surveys are self-administered whereas face-to-face surveys are interviewer-administered. This comes with advantages as well as disadvantages for both modes and with positive as well as negative effects on data quality. On the one hand, an interviewer is able to motivate people to participate and thereby increase response rates (Groves et al. 2004), and the interviewer can support the respondent with questions requiring a high cognitive effort (Holbrook, Green, and Krosnick 2003). On the other hand, the presence of an interviewer increases the normative pressure on respondents to provide socially acceptable answers, whereas the anonymous environment of self-administered interviews allows them to be more honest (Tourangeau and Smith 1996; Tourangeau and Yan 2007). Especially with regard to more sensitive questions, the mode of data collection has a strong influence on biases due to social desirability (Chang and Krosnick 2010; Tourangeau and Yan 2007). Research has shown that factual measurements of sociodemographic characteristics are less affected by the interview situation, whereas subjective and private questions, which score higher on sensitivity, elicit a stronger mode effect, depending on whether they are self-administered or interviewer-administered (Burkill et al. 2016; Christensen et al. 2013). In terms of content that is important for family research, previous findings on social desirability bias between modes have often concentrated on traditional or obvious sensitive items that are strongly normative, such as attitudes toward gender and family

roles (Liu 2017; Liu and Wang 2016), questions regarding sexual behavior and sexual experiences (Burkill et al. 2016; Kelly et al. 2013), and questions regarding mental and physical health (Braekman et al. 2020; Christensen et al. 2013).

The question remains open as to the extent other kinds of subjective questions are perceived as sensitive by respondents, might therefore be prone to biases based on social desirability, and thus might also be subject to effects based on the mode of data collection. The research field of relationship quality provides a very good example for analyzing this question. First of all, it is a frequently addressed topic in research about intimate relationships, with a high relevance for other family-related events, such as childbearing (Rijken and Liefbroer 2009) and union dissolution (Karney and Bradbury 1995). It might also affect personal matters, such as well-being (Gustavson et al. 2013). Second, the experimental study we use has many suitable indicators for measuring relationship quality comprehensively. These indicators cover a broad range of domains in a relationship as we study perceived relationship stability, satisfaction with different aspects of couples' daily life, different areas of conflict, and different levels of aggressive and violating conflict styles. Furthermore, these indicators provide good examples of subjective perceptions for which we lack clarification as to what extent they must be considered sensitive questions. One can assume that the normative expectation regarding the maintenance of a happy relationship in studies about intimate relationships is strong, which in turn increases respondents' perception of social pressure in a personal interview situation. The measurement of relationship quality requires subjective assessments by the respondent and may also be perceived as sensitive, depending on the individual situation. This may be true for at least some of the aspects of relationship quality for which we find indicators in our dataset. The broad spectrum of domains of relationship quality covered by the data may even provide a nuanced picture. Therefore we investigate whether measurements of this construct differ between a self-administered and an interviewer-administered interview situation.

For assessing differences between modes of conduction, an experimental pilot study was carried out in Germany within the Generations and Gender Programme (GGP), which compares the traditional face-to-face mode with the upcoming web mode. The GGP is a well-established, large-scale survey program in family research. We profit from a unique experimental setting that allows us to use an existing survey instrument, the GGS, as well as an experimental design to test for differences in the mode of conduction. The aims of this study are to examine whether differences in measurements of frequently used indicators of relationship quality occur between face-to-face and web modes and to assess which survey design provides the most reliable measurement of relationship quality. Our research question is therefore: Do measurements of indicators frequently used for explaining relationship quality conducted in face-to-face mode differ from measurements conducted in web mode? In a first descriptive step, we compare mode-

specific differences between the distributions of the particular indicators for relationship quality in the two experimental groups. In a second step, we estimate multivariate regression models to adjust for family demographic confounders to assess the impact of mode on the particular items of relationship quality. Given that in each mode, persons with specific characteristics might be more or less likely to participate, the regression models allow us to control for such selective confounders.

In the context of continuing methodological innovations and developments in data collection, this study should sensitize data users to the possibility of distortions for frequently used key variables in substantive analysis due to survey mode decisions. The findings are especially relevant for data analysis as well as for data conduction of cross-national and panel surveys based on different modes of data collection.

2. Background and expectations

2.1 Face-to-face versus web interviewing

Face-to-face and web surveys differ mostly regarding the degree of interviewer involvement. According to Couper (2011), this has an impact on overall participation in a survey, thus on the response rate and data quality. A meta-analysis revealed that response rates of web surveys are lower compared to more traditional modes of data collection, such as face-to-face surveys (Daikeler, Bošnjak, and Manfreda 2020). But the last years have also shown that, at least in Western European countries, response rates of data collections in face-to-face mode have declined rapidly (Beullens et al. 2018). Nevertheless, interviewers can be helpful in the recruitment stage in motivating the target person to participate. Furthermore, web surveys obviously bear the risk of underrepresenting the off-line population (Schonlau et al. 2009). This is particularly relevant for surveys that have to rely on nonprobability samples – due to lack of a suitable sampling frame for the particular target population, for example, which lowers the representativeness needed for large-scale social science surveys (Tourangeau 2017). Moreover, the interviewer can play a helpful role during the interview by assisting the respondent in the response process. Interviewers are able to support the respondent in answering questions requiring a high cognitive effort. They can motivate and support the respondent verbally as well as through nonverbal communication throughout a long interview (Holbrook, Green, and Krosnick 2003). Some studies showed that self-administered web surveys had higher proportions of item non-response, higher proportions of choosing “Don’t know” answers, and less differentiation on rating scales than face-to-face surveys (Heerwegh 2009; Heerwegh and Loosfeldt 2008).

However, data collectors have to consider that the use of interviewers is considerably more expensive than conducting a web interview. Additionally, there are also positive effects on data quality resulting from the absence of an interviewer. While in face-to-face interviews, interviewers have the locus of control over the whole interview process, in web interviews, respondents have the autonomy to answer the questionnaire at the time and place they prefer, at the speed that suits them best, and with the option to stop during the interview and continue later on (Couper 2011). Furthermore, web interviews are characterized by a higher degree of privacy and anonymity than face-to-face interviews. This may be expected to modulate the strength of social desirability effects, as discussed in detail below.

2.2 Social desirability bias and relationship quality

Social desirability explains most prominently why the mode of data collection plays such an important role in the answering of questions prone to sensitivity. According to the concept of social desirability, respondents tend to overreport socially desirable answers and underreport socially undesirable answers (Callegaro 2008). An open question is how strongly particular questions are affected by social desirability bias. Often, the strength of social desirability bias corresponds with the degree of sensitivity of the question (Krumpal 2013). But the perceived sensitivity of a question depends strongly on the person who is interviewed and thus also on his or her individual situation and how much emotional stress the respondent would endure by giving an honest answer. This can further vary by cultural, social, and situational context (Lee and Renzetti 1990). Given that these factors vary across and even within studies due to different questionnaire contents and target populations, it is difficult to draw general conclusions about social desirability bias.

Taking the mode of conducting surveys into account, research has shown that respondents tend to answer more truthfully and honestly in an anonymous interview situation, especially for obviously sensitive questions (Chang and Krosnick 2010; Tourangeau and Smith 1996; Tourangeau and Yan 2007; for an overview: Krumpal 2013). In other words, interviewer-administered modes, such as face-to-face mode, lead more often to socially desirable responding because respondents tend to present themselves in a socially favorable manner instead of reflecting the true situation in front of the interviewer. In contrast, in self-administered interviews, such as web surveys, respondents have a higher level of privacy and a lower level of perceived social pressure and show more honest answer behavior (Heerwegh 2009). Interviewer characteristics, such as gender or ethnicity, can also affect interview dynamics and impact responses in personal interviews, particularly for questions prone to social desirability biases and for

questions related to these characteristics (Davis et al. 2010). For example, research shows that the gender of the interviewer can influence response behavior regarding marriage-related questions (Liu and Stainback 2013). Moreover, the degree of familiarity between the respondent and the interviewer might also impact the respondent's effort in answering sensitive questions, as respondents show lower levels of trust and disclosure when the interviewer is a stranger and is not familiar with the local environment (Weinreb, Sana, and Stecklov 2018).

To reduce bias related to social desirability in face-to-face interviews, highly sensitive questions are often surveyed in a computer-assisted self-interview (CASI) module, where the respondent can complete individual question blocks independently. Even though this is a good way to make respondents' answers more anonymous, the control still remains with the interviewer. By contrast, in an entirely self-administered mode, such as web, the respondent can show a higher degree of self-disclosure because there is no other person present and therefore no time pressure and a free choice of where to respond to the interview. Additionally, face-to-face interviews are mostly conducted in the respondent's own household, which means respondents might be influenced not only by the interviewer but also by any additional persons present, such as partners, spouses, or children (Schröder and Schmiedeberg 2020). So-called bystander effects have an impact on answering questions in computer-assisted personal interviews (CAPI) as well as in CASI during face-to-face interviews; research shows that the reporting of less desirable answers in web mode has the highest response accuracy compared to other self-administered modes (Kreuter, Presser, and Tourangeau 2008).

Previous research on mode-related social desirability bias in studies about families and intimate relationships concentrates on a few subjective and objective indicators, but as far as we know, no study has investigated the impact of social desirability bias on items about relationship quality. An early experiment from the National Survey of Family Growth discloses a higher reported number of abortions in self-reports than in interviewer-administered interviews (Fu et al. 1998). This is one example of highly sensitive information in family research. A more recent experiment of the third British National Survey of Sexual Attitudes and Lifestyles examined changes in responses from the same respondents in interviews first conducted in CAPI and CASI and afterward conducted in web mode. The findings show that not all sensitive questions regarding sexual life revealed a mode effect between CAPI, CASI, and web. But for some questions regarding individual behavior, such as same-sex experiences, and for opinion questions, such as those about sexual satisfaction, the study found a higher level of self-disclosure and more socially undesirable answers in web interviews compared to CAPI and CASI (Burkill et al. 2016), which implies that even a switch to CASI mode cannot fully compensate for the downsides of conducting personal interviews. The anonymous interview situation can play a major role here, as Robertson and colleagues (2018) find

that respondents in online surveys report the highest comfort level in answering questions about non-heterosexual prevalence compared to 16 other interviewer-administered as well as self-administered survey modes. An explanation is that online interviews are perceived as less intrusive and as having a higher level of anonymity and privacy without creating the feeling of being observed or recorded (Robertson et al. 2018). In the field of public health research, experiments come to similar findings regarding opinion questions and subjective assessments of personal well-being and health status. Answers to factual questions are comparable between face-to-face and self-administered survey modes, but they detect different levels of mode effects in the answering of more sensitive questions involving subjective assessments (Braekman et al. 2020; Christensen et al. 2013).

As prior research shows, mode effects can vary strongly between studies, because every survey focuses on different topics, uses different questionnaires, and aims at different target populations. This means effects between face-to-face and web surveys regarding biases due to social desirability are hard to generalize (Couper 2011). Therefore it is always necessary to evaluate such effects in the context of the particular study. What also can come into play in surveys about intimate relationships is a social expectation about how a relationship should ideally be. In such surveys, respondents might feel embarrassed to admit that their own situation does not conform to a norm or expectation about happy relationships. This means that perceived social pressure might lead to misreported feelings or subjective assessments (DeMaio 1984). What underlines this assumption is also a selective trend in the general participation in studies about families and intimate relationships: People with happier and closer relationships within a family are more likely to participate in such surveys (Kalmijn 2021).

Items that assess relationship quality are subjective questions that require respondents to reflect on their behavior and feelings. Our indicators under study serve to assess the quality of an intimate relationship. However, they cover a broad range of domains in an intimate relationship, such as household task division, child care, feelings and doubts about the relationship, and ways of dealing with conflict. These different topics can seem more or less sensitive for the respondent, depending on the individual situation, especially when they touch on a sore point in a relationship. Due to the higher anonymity, we expect that web interviews support more open and probably more honest answer behavior. Further, the locus of control is up to the respondent, which may lead to a higher comfort level in answering unpleasant private questions. This means that the respondent can fill out the questionnaire at the time and place of their own choice – for example, when they are alone at home, so that interference by a third person, such as a partner, can be avoided. Therefore we expect a mode effect for indicators of relationship quality as follows: Respondents who participate in web mode provide more socially undesirable answers and higher levels of self-disclosure than face-to-face respondents. This means that web respondents should report a lower level of relationship quality and

assess the relationship on average more negatively compared to respondents in a face-to-face interview. Nevertheless, it is an open question as to what extent indicators on relationship quality display such a mode effect and to what extent they must be considered to be sensitive questions in the context of family research.

3. Data and methods

3.1 Experimental design and case selection

For answering our research question, we used data from an experimental pilot study within the framework of the GGP (Emery et al. 2018). The GGP is an international family demographic infrastructure that conducts the GGS. The GGS is fielded in many European and a few non-European countries and is designed as a three-wave panel study. The study focuses on families, intimate relationships, and life course trajectories of individuals (Gauthier, Cabaço, and Emery 2018). The GGP pilot study was conducted in Germany, in addition to two other countries, in 2018. The aim of the pilot study was to test whether a revised version of the GGS questionnaire and a new survey design work well in the field and whether the GGS can be conducted as a mixed-mode or online survey. A push-to-web design was applied and compared with the traditional GGS mode, which is face-to-face mode (Lugtig et al. 2022). Push-to-web design means that we conducted web interviews as we would have done in an entirely online survey but contacted non-respondents again after the web fielding period and asked them to participate in a personal interview. For our research question, we concentrate on the web respondents of that group. The German pilot study carried out further experiments regarding the timing and amount of incentives. As variation in incentives may affect data quality and response behavior regarding sensitive questions (Medway 2012), we compared only groups that used identical incentives. Only in this way could we obtain an experimental setting that provided the same initial conditions for both groups, except the mode of conducting the interview, which serves as the treatment.

Respondents in the reference group participated in a CAPI, and respondents in the experimental group participated in a computer-assisted web interview (CAWI). Both the face-to-face group and the web group received the same Blaise-programmed GGS questionnaire in terms of question wording, routing, and design. Further, both groups received a prepaid incentive worth five euros. The target persons, aged 18–49, were selected with simple random sampling from local registry offices (Einwohnermeldeämter) in the German federal state of Bavaria, with a quota of 50% of addresses coming from rural areas and 50% of addresses coming from urban areas. The target persons were randomly assigned to the experimental groups. The size of the gross

sample was calculated by the fieldwork institute on the basis of the expected response rate per mode. The aim was to achieve at least 200 cases per experimental group. Based on experience with other German surveys, the gross sample size of the face-to-face group was set lower because the response rate for the face-to-face mode was expected to be higher than for the web mode.

Both groups received an invitation letter with the unconditional incentive in it. For the face-to-face group, the letter announced that an interviewer would come to the household to conduct the interview. For the web group, a URL with a password was provided in the letter, and target persons were asked to go online and fill out the questionnaire on their own. The web group received also two reminder letters, each two weeks after the previous letter. Table 1 gives an overview of the design specifications and the case selection. The overall response rate was calculated according to Response Rate 1 following the AAPOR classification of standard definitions (AAPOR 2016). The response rate showed that we had a higher participation – by nearly 10 percentage points – in the face-to-face group. Both response rates are rather low, but other German social science surveys conducted face-to-face yield similar response rates (Wolf et al. 2021).

Table 1: Overview of the experimental design and case selection

	Reference group	Experimental group
Mode of data collection	Face-to-face (CAPI)	Web (CAWI)
Country where conducted	Germany	Germany
Target population	18–49 years old	18–49 years old
Incentives	Five prepaid euros	Five prepaid euros
Maximum number of contacts	Invitation letter + five personal contact attempts	Invitation letter + two reminder letters
Gross sample	685	1,365
Net sample	193	261
Response rate in %	29.5	19.4
Respondents with a partner	146	197
Respondents with a partner in %	76.0	77.6

Source: GGP pilot study 2018; authors' own calculations.

Our research question focuses on the assessment of relationship quality in couples, so we only considered respondents in our analyses who reported that they had had a partner for at least three months. That is how the GGS measures partnership status. Only these respondents could actually answer questions regarding their current relationship. Table 1 gives an overview of the number of cases in the sample under study. It can be seen that the proportion of people in a relationship does not differ greatly between the two groups, with 76.0% of respondents having a partner in face-to-face mode and 77.6% of respondents having a partner in web mode. This corresponds closely to the proportion of persons with partners in other German studies about families and intimate relationships (Kantar Public 2018). By comparing the distribution of family demographic characteristics in the overall sample and our analytical sample, including only

respondents with a partner, for each mode shown in Table 2, we see that slightly more women and fewer younger people have a partner. However, these tendencies are evident for both modes and can therefore be ignored. Generally, respondents in the web sample – irrespective of having a partner – are higher educated and more likely to live in urban areas compared to face-to-face respondents (see Table 2), which is consistent with existing research (e.g., Atkeson, Adams, and Alvarez 2014).

Table 2: Family demographic distributions (in percent) for all respondents in the entire sample and for only respondents with a partner in the sample under study

	Face-to-face		Web	
	All respondents	Only respondents with a partner	All respondents	Only respondents with a partner
N (observations)	192	146	254	197
Sex				
Male	47.92	42.47	47.24	45.69
Female	52.08	57.53	52.76	54.31
Education				
Low	25.52	21.92	12.90	14.66
Middle	26.04	28.08	33.06	35.60
High	48.44	50.00	54.03	49.74
Age				
18–29	37.70	25.52	30.68	23.71
30–39	29.32	35.86	32.27	32.99
40–49	32.98	38.62	37.05	43.30
Citizenship				
German	88.02	87.67	88.54	87.76
Non-German	11.98	12.33	11.56	12.24
Regional setting				
Urban	39.58	41.10	54.33	51.78
Rural	60.42	58.90	45.67	48.22
Child under 6 in household				
No	60.94	51.37	66.93	57.87
Yes	39.06	48.63	33.07	42.13
Relationship status				
Married	62.33	62.33	69.79	69.79
Cohabiting	24.66	24.66	17.19	17.19
Living apart together	13.01	13.01	13.02	13.02

Source: GGP pilot study 2018, authors' own calculations.

3.2 Methodological approach and measurements

3.2.1 Methodological approach

The design of the experimental study allows us to compare answering patterns in face-to-face mode and web mode to assess the overall impact of one mode compared to the other. We analyze 15 single items in the univariate analyses and six items in the multivariate regression analyses. All indicators under study relate to the construct of relationship quality.

In a first step, we apply a univariate approach and calculate means and proportions on item level to see how the mode affects point estimators. We test for mode-specific differences in the distributions with a two-sample t-test for mean differences and a Pearson-χ^2-test for independence between categorical variables. Univariate comparisons should reveal initial ad hoc findings regarding the extent of distortion due to social desirability and self-disclosure in one mode as compared to the other. Further, surveys often use ex-post weighting to adjust for certain biases due to mode-specific selectivity, non-response bias, or coverage bias (Groves et al. 2004). Hence we additionally apply post-stratification weighting to evaluate whether the measurement equivalence between the univariate distributions of the two modes improves or not (Bethlehem and Stoop 2007; Schonlau and Couper 2017). For example, if more young people participate in web than in face-to-face mode, this may reduce the average duration of relationships in one mode, which might have a confounding effect on relationship quality. Therefore we adjust the entire sample according to population totals for specific demographic characteristics that are available from official German statistics (census and micro census) for our target population. As auxiliary variables for weighting, we use sex, age groups, highest level of education, nationality, and regional setting. Selective non-response can have many sources and might not be based exclusively on demographic characteristics (Schonlau, van Soest, and Kapteyn 2007), but we had to rely on the best information available from official statistics, which include only demographic information for our target population. A detailed list and the sources of information used for post-stratification weighting can be found in Table A-1 in the appendix.

In a second step, we pool the experimental and the reference group and apply multivariate regression analysis with mode as the explaining variable and a block-wise adjustment of further confounding variables to test whether the effect of mode is robust or not. Given that participation in surveys might be selective, regression analysis allows us to control for characteristics that correlate with selective participation in one mode. Indicators of relationship quality are treated as outcome variables and the mode of conducting the survey as a predictor variable. In the first baseline model, we estimate the crude mode effect on the respective indicator of relationship quality. In the second model, we include the same standard demographic variables we used for post-stratification

weighting to adjust for selective participation. In the third model, we include additional family-related variables, which are often used as adjustment variables for relationship quality. We then examine whether a possible effect of the mode of data collection on the respective indicators of relationship quality changes between the models or whether the effect is robust after adjusting for further explanatory determinants. For the one binary dependent variable, which is the question whether or not the respondent had thought about breaking up, we calculate the linear probability model (LPM) because it facilitates the interpretation of estimates, especially when comparing coefficients across differently specified models. As a robustness check, we further apply logistic regression models and estimate average marginal effects (AME) that yield results similar to those from the linear probability approach (see Table A-4 in the appendix). For the other indicators, we perform linear ordinary least squares (OLS) regression models and show the estimated coefficients.

3.2.2 Measurements

The items evaluated in our study are often used as predictors, mediators, and outcomes in substantive analyses in research about the quality of intimate relationships. They are included in many large-scale surveys about families and relationships, such as the GGS.

In the univariate approach, we examine 15 indicators of relationship quality separately to get an impression of mode effects on a broad variety of items that potentially display different effects. One item relates to subjective stability, three items relate to satisfaction, and the remaining 11 items relate to conflict frequencies and styles. Starting with subjective instability, we evaluate the question of whether the respondent has thought about breaking up with their partner. The binary indicator is coded with 0 for no and 1 for yes. Items on satisfaction are an often-used survey instrument to assess feelings and are mostly measured on point scales. The GGS questionnaire contains three satisfaction scales, which cover three different domains in intimate relationships. The question wording is: "How satisfied are you with the relationship in general, the division of household tasks, and the division of child care tasks." The answers have to be assessed separately on an 11-point scale, where 0 means very dissatisfied and 10 means very satisfied. A rating of 5 means medium satisfaction. The question on satisfaction with the division of household tasks was asked only to respondents who have a coresidential partner, and the question regarding satisfaction with the division of child care tasks was filtered for parents. All other questions on the various aspects of relationship quality, including those on conflicts, were asked to all respondents who have a partner, regardless of other criteria.

Whereas the first four indicators relate more to the level of feelings, the last indicators concern behavior in a relationship, specifically conflict behavior within the couple, differentiated into frequency of conflicts and conflict styles. For the univariate approach, we use seven single items regarding the frequency of conflicts on the following issues: household chores, money, leisure time, relations with friends, relations with parents, having children, and child-raising. The answer categories range from 1 (never) to 5 (very frequently). For the multivariate analysis, we generate one indicator for the frequency of conflicts within the couple in general, summarizing the information of these seven items to reduce complexity. Given that each single item measures the frequency of certain conflicts on the same scale, we are able to directly compare answer codes. Assuming that one relationship conflict will rarely touch several of the issues represented by the seven items at the same time and that the conflicts measured by the seven items have little overlap, we consider the addition of answer codes as an appropriate way of constructing such an indicator for frequency of conflict. Accordingly, we generate an additive index (Cronbach's α: .70), which ranges from 1 (no reported conflicts) to 29 (very frequently reported conflicts). It is recoded such that the more conflicts reported, the higher the value of the index.

Finally, we look at four single items that cover reactions in conflict situations. These items assess how often respondents avoid discussions by giving in, discuss conflicts calmly, argue heatedly or get loud, or refuse to talk. Here again, answers are coded from 1 (never) to 5 (very frequently). These conflict styles are also summarized into one indicator to reduce complexity for multivariate analyses. The indicator measures the tendency of choosing inadequate conflict behavior according to the social norms of a late-modern society in which it is expected that disagreements are resolved by rational exchange of arguments. Accordingly, we recode the item "discuss conflicts calmly" reversely and construct an additive index summing up all four items, which are all recoded so that the more inadequate the conflict behavior reported, the higher the score (Cronbach's α: .54). We are aware that, in this case, we are summarizing information from more heterogeneous items, so that the validity of the generated indicator is lower. Although Cronbach's α of the second index has a lower internal consistency, the scale is sufficient for our purposes to get an additive measure of inadequate conflict behavior. The index ranges from 1 (only inadequate conflict behavior reported) to 17 (only adequate conflict behavior reported). The original wording of all questions and answer categories can be found in Table A-2 in the appendix.

As control variables for the multivariate approach, we use the same demographic indicators as for the post-stratification weighting. Sex of a respondent is coded as 0 for male and 1 for female. For nationality we distinguish between 0 for German citizenship and 1 for non-German citizenship. Age is measured in years and ranges from 18 to 49. Education is measured dichotomously: 1 for highest education (college entry

qualification) and 0 for lower/middle school education or less. Information on regional setting or community size was provided by the fieldwork institute and divides areas where respondents live into (1) urban areas and (2) rural areas. These background variables are not only used frequently as standard demographic controls; they can also affect participation in web mode (Vehovar et al. 2002). To control for family-related determinants, which often correlate with relationship quality, we include the variables relationship status and the existence of coresident children under age 6. We decided to choose an indicator for having children of younger ages because we assume that infants affect couples' daily lives more than older children, as they need more care and attention, often at the expense of the young parents' relationship quality and time. Additionally, parents of younger children have a lower level of mobility and available time, which can impact participation in the respective survey mode. As a sensitivity check, we also calculate models using an indicator for having children of any age as a control variable, and we find no differences in the identified mode effects. The measurement of relationship status distinguishes between respondents who are married to their current partner, irrespectively of cohabitation (1); respondents who live together with their partner without being married (2); and respondents who have a partner but are not living in cohabitation or marriage with that partner (living apart together; 3), often in long-distance relationships. The other indicator distinguishes between "at least one child under six years living most of the time in the same household with the respondent" (1) versus "no children under six years living in the same household with the respondent" (0). Unfortunately, information about the duration of the current relationship is unavailable in the GGS.

4. Results

4.1 Univariate analyses

We start with a look at the univariate distributions of the indicators of relationship quality in the two experimental groups to examine whether the mode of conduction affects point estimators under the two experimental conditions. The distributions for all 15 single indicators of interest are shown in Table 3. Respondents using the web report higher shares of socially undesirable answers than do face-to-face respondents for almost all items under study.

Table 3: Means of indicators or percentage of confirmative answers by mode of data collection with 95% confidence intervals

	F2F	Web	Mode difference	F2F	Web
	Mean or percentage		Δ	n	
Subjective instability [a]					
Thought about breaking up (in %)	8.97 (5.27–14.86)	17.13 (12.30–23.35)	**8.16**	145	181
Satisfaction [b]					
General relationship	9.11 (8.91–9.31)	8.63 (8.43–8.82)	**–0.48**	145	190
Household tasks	8.46 (8.19–8.73)	7.90 (7.59–8.20)	**–0.56**	127	166
Child care tasks	8.66 (8.28–9.03)	8.11 (7.73–8.50)	**–0.55**	70	81
Conflict frequency [c]					
Household chores	2.31 (2.13–2.48)	2.45 (2.31–2.59)	**0.14**	146	196
Money	1.66 (1.53–1.79)	1.97 (1.83–2.12)	**0.31**	146	195
Leisure time	2.23 (2.07–2.38)	2.39 (2.26–2.51)	**0.16**	146	196
Relations with friends	1.50 (1.39–1.61)	1.69 (1.57–1.81)	**0.19**	145	196
Relations with parents	1.67 (1.53–1.81)	1.85 (1.72–1.99)	**0.18**	146	195
Having children	1.30 (1.17–1.42)	1.28 (1.18–1.38)	**–0.02**	145	192
Child-raising issues	1.80 (1.64–1.96)	1.94 (1.80–2.08)	**0.14**	142	182
Conflict style [d]					
Avoid discussion by giving in	2.61 (2.45–2.77)	2.66 (2.53–2.79)	**0.05**	132	184
Discuss conflicts calmly	4.06 (3.91–4.21)	3.86 (3.72–4.00)	**–0.19**	136	180
Argue heatedly or get loud	1.88 (1.75–2.02)	2.13 (2.00–2.25)	**0.25**	136	186
Refuse to talk	1.74 (1.57–1.90)	1.88 (1.75–2.01)	**0.14**	136	183

[a] Reference category is "not thought about breakup."
[b] From 0 (very dissatisfied) to 10 (very satisfied).
[c] From 1 (never) to 5 (very frequently).
[d] From 1 (never) to 5 (very frequently).
Notes: F2F = face-to-face; Δ = mode difference (in bold); 95% confidence intervals in parentheses.
Source: GGP pilot study 2018; authors' own calculations.

When we take the different content dimensions of the indicators into account, we can see that especially those items that concern feelings, such as satisfaction and the perceived stability of a relationship, display stronger mode differences. Based on these univariate findings, more than 17% of web respondents – nearly twice as many respondents as in face-to-face mode, with 9% – confirm that they have thought about breaking up with their current partner. Correspondingly, web respondents rate their general relationship satisfaction nearly 0.5 points lower on an 11-point-scale than do face-to-face respondents. The same is true for satisfaction with daily routines in a relationship, like the division of child care and household tasks. For the content-specific frequencies

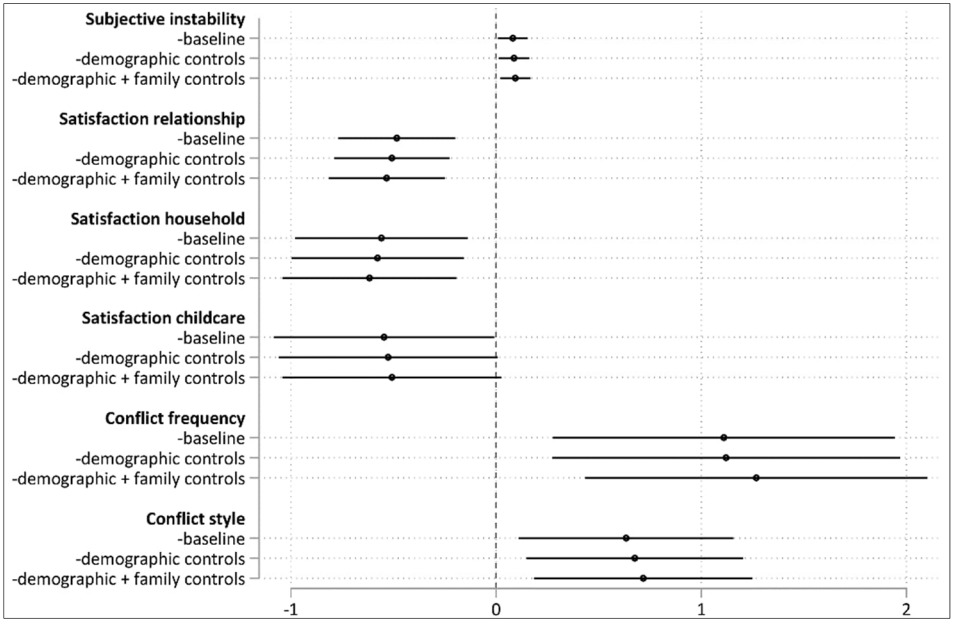
of conflicts as well as for the different conflict styles, most items show mode differences, with a higher reported frequency of conflicts and inappropriate conflict behavior in online interviews compared to face-to-face, but with varying magnitudes of mode differences between the single conflict items. The item on the frequency of conflicts regarding having children shows a very small and therefore negligible mode difference in the opposite direction. One explanation might be that a majority of the respondents already have children, so this topic was not leading to conflicts between parents anymore. The rather aggressive conflict behavior "argue heatedly or get loud" shows larger mode differences, whereas the comparably modest conflict style "avoid discussion by giving in" reveals almost no difference between the modes. Because avoiding a discussion might not be a socially undesirable way of dealing with your partner, this conflict behavior is not as clearly indicative of a bad conflict style as the others.

With the help of post-stratification weighting by adjusting sample distributions to the reference distributions of our target sample, we try to control for biases due to selective participation. The weighted distributions of the indicators of relationship quality are very similar, however (see Table A-3 in the Appendix). This emphasizes that ex-post weighting by the demographic indicators for which reference data are available cannot adjust for the mode-specific differences for our items under study.

4.2 Multivariate analyses

We continue to use unweighted data for our analysis, as mode differences were the same for unweighted data and for ex-post weighted data. We estimate three models for each of our six outcome variables: a baseline model without control variables, a second model with our demographic confounder variables, and a third model with all demographic and family-related control variables. Figure 1 displays the effect of web mode on each indicator of relationship quality separately compared to the reference face-to-face mode. Because the focus of this study lies in the evaluation of the mode effect, we refrain from showing the regression results of the control variables. The regression tables of the mode effects can be found in Table A-4 (see Appendix). As described in the methodology section, the single items for frequency of conflict and conflict styles are summed up to two indexes. For each indicator, the effect of web mode is shown – first as a single effect in a baseline model; second adjusted by standard demographic variables, which should control for selective participation; and third under additional adjustment of family-related confounders. The size of mode effects can be compared only across models of the same outcome variable, not between different outcome variables.

Figure 1: Effect of web mode on indicators of relationship quality with 95% confidence intervals under block-wise adjustment of demographic and family-related control variables



Notes: Only the coefficients for the effect of web mode (compared to face-to-face mode), as the main explaining variable, with 95% confidence intervals, are shown. Coefficients of binary indicator subjective instability are estimated with linear probability models; coefficients of all other indicators are estimated with linear regression models. Demographic controls are sex, nationality, age, education, and regional setting; family-related controls are relationship status and coresident children under 6.
Source: GGP pilot study 2018; authors' own calculations.

The multivariate findings in Figure 1 confirm that the effect of web mode is robust for all indicators on relationship quality even when we adjust for demographic and family-related variables. By comparing the effect of web mode across the three different models for one indicator, we see that the effect is either stable across the models or slightly increases. Therefore we focus on reporting the findings based on the third model, including demographic and family-related control variables.

Starting with the indicator on subjective instability, the estimated coefficients based on LPM show that respondents in web mode have a probability of reporting that they thought about a separation that is 9.4 percentage points higher than the probability for respondents who were asked this question in a face-to-face interview. This means that respondents in the more anonymous web mode are more likely to report that they have thought about breaking up with their current partner than respondents interviewed in a

personal interview. The second indicator frequently used for measuring general relationship quality is overall satisfaction with the relationship. The results confirm the univariate findings: Even under control of demographic and family-related variables, respondents in web mode rate their satisfaction with the relationship about 0.5 points lower on an 11-point scale than respondents in face-to-face interviews. The same pattern can be seen for the reporting of satisfaction with specific domains, such as household tasks and child care. Respondents in web mode assess their satisfaction with household tasks more than 0.6 scale points lower than respondents in face-to-face mode. Results are similar for the assessment of satisfaction with child care tasks: Online respondents rate their satisfaction about 0.5 scale points lower on a scale from 0 to 10 than face-to-face respondents. As can be seen in Figure 1, the confidence interval of the mode effect on satisfaction with child care touches the zero line slightly, which might be explained by the low number of persons who answered this question, as this item was posed only to respondents who have children.

Apparently, feelings and thoughts about the relationship are assessed more negatively in web mode than in face-to-face mode, which speaks for a higher level of self-disclosure and less socially desirable answers in web surveys. In other words, the findings support the assumption that respondents in web interviews are more likely to report that they are less satisfied with their current relationship and are more likely to doubt the stability of the relationship.

The last two indicators focus on the assessment of behavior. For the indicators frequency of conflicts and conflict styles, we find a higher reporting of conflicts and inappropriate conflict behavior in web mode compared to face-to-face mode. The reporting of the number of conflicts on various topics increases by 1.3 points on a scale from 1 to 29 when respondents answer in web mode compared to face-to-face mode. The effect is similar for the reporting of inappropriate conflict behavior: Compared to respondents in a personal interview, respondents who participate in a web survey report a 0.7 scale points higher level on a scale from 1 to 17.

In summary, regarding the reporting of feelings as well as behavior, web respondents show a consistently higher socially undesirable response behavior than respondents in face-to-face interviews. Considering all indicators of relationship quality examined in this study, respondents in web mode assess their relationships more negatively than those in face-to-face mode.

5. Discussion

Our analyses use experimental survey data to assess the existence and the extent of a mode effect when comparing two particularly different modes of data collection, web

mode and face-to-face mode, on measurements of relationship quality in surveys about families and relationships. Our findings show clear differences for almost all indicators that assess various aspects of the quality of intimate relationships between respondents interviewed in a traditional face-to-face design and respondents who participate in self-administered web interviews. Web respondents are more likely to state that they thought about breaking up. They assess a lower relationship satisfaction in general as well as with respect to the distribution of household chores and child care responsibilities. And they report more conflicts in their partnership as well as higher shares of aggressive or non-constructive conflict behavior. These indicators not only cover different content-related aspects in the context of intimate relationships, but they also rely on a broad range of subjective assessments, such as feelings, behavioral patterns, and experiences. All in all, respondents who participated in web mode report a lower quality in intimate relationships. These effects are robust, as they remain stable after controlling for demographic and family-related confounders that correlate with relationship quality and survey participation.

Our results support the assumption that the anonymous and private interview situation of web surveys, compared to traditional face-to-face surveys, leads to a smaller subjectively perceived exposure to social desirability, thereby impacts the responses of interviewees, reduces bias due to social desirability responding, and thus improves the validity of measurements. According to our expectations, web respondents give more socially and normatively undesirable answers and report a less rosy picture of their partnership life than do face-to-face respondents. Further, the findings could indicate that respondents who participate online have a higher willingness to self-disclose than respondents who are confronted with an interviewer, which is in line with existing research (Robertson et al. 2018; Burkill et al. 2016). One could assume that measurements on relationship quality conducted in web surveys show a more realistic picture of today's couple relationships than those conducted face-to-face.

The findings further indicate that the assessment of relationship quality must be considered as highly sensitive and generally biased by effects of social desirability – in web mode to a lesser degree than in an interviewer-administered mode. However, we cannot prove for a general underreporting of sensitive behavior as we have no reference value of the real situation and can assess only differences in answer behavior between two modes of data collection. This result of relationship quality being a sensitive topic in surveys is relevant in particular for studies about intimate relationships, because surveys in this context are, for the same reason, confronted with the risk of selection biases toward happier and closer relationships (Kalmijn 2021). One can assume that respondents who are actually less satisfied with their relationship and unhappy with their partner tend to be generally underrepresented in a family survey and are therefore of particular interest.

Depending on the individual situation, some aspects in an intimate relationship might score higher on sensitivity than others and cause gradually stronger biases based on social desirability. As shown in our descriptive findings, single items on conflictual behavior differ in their magnitude of mode differences, which might indicate that the according behaviors are perceived as differently strongly undesirable. For example, refusal to talk may be less undesirable than aggressive and potentially threatening conflict behavior. At the same time, we can assume that similar mode effects would be found for most subjective perceptions and evaluations in other research topics within family demography and beyond; many such indicators might be perceived as sensitive by respondents, as they may expect the interviewer or others to have certain opinions and according expectations regarding an acceptable answer. The more plausible that regarding a certain subjective question, a social norm exists, the more likely it is that such an indicator will be biased by effects of social desirability and affected by mode effects.

In our study, we use two extremes of interviewer involvement as an experimental design to sensitize primary researchers as well as data users about the impact of a design decision on data. Nevertheless, there are also mixed-mode designs or hybrid modes of conducting interviews that can be placed on a gradient between face-to-face mode and web mode, such as CASI modules applied within a personal questionnaire, and these could improve measurement equivalence. A limitation of this study is therefore that it remains an open question as to whether the anonymous setting of the web is decisive for the higher degree of disclosure and lower degree of social desirability or whether the presence of an interviewer or other bystanders might be compensated for by a CASI switch. Even if one could assume less socially desirable answer behavior in CASI than in a face-to-face interview, experimental studies show that web interviews reveal the highest degree of self-disclosure for sensitive questions compared to other self-administered modes (Burkill et al. 2016; Kreuter, Presser, and Tourangeau 2008). Nevertheless, CASI switches should be used more frequently in personal interviews with intimate and subjective questions, such as those on relationship quality, when a web interview is not possible.

Another limitation is that our experimental study relied on a small number of cases due to budget constraints, as is the case for most experimental studies. Due to the low number of observations, detailed analyses with subgroups differentiated by gender or age could not be carried out. Thus the methodological approach remained limited. It would be imaginable that, for example, women would be less affected by biases due to social desirability and mode effects than men, since women generally report lower relationship qualities and tend to break up relationships more often than men do. It would be imaginable that parents may be more affected by mode effects than people in childless relationships since maintaining a stable relationship may be more strongly socially desired if a child is involved. However, such assumptions require further investigation.

While the results in detail must thus be interpreted with caution, the mode effects nevertheless proved robust and revealed a stable pattern across several indicators, which allows us to consider our main findings reliable. It would be highly valuable for family research to analyze whether the measurement of the impact of relationship quality on substantive outcomes, such as breakups or divorces, is also affected by the mode. Unfortunately, this could not be tested in our study due to small case numbers and lack of a longitudinal design.

We conclude that data users should be aware of the need to control for the mode of data collection when analyzing data on relationship quality collected in different modes, especially when self-administered as well as interviewer-administered modes are involved. It is important not only to assess data for the representativeness of sociodemographic indicators and, if necessary, weight the data and adjust for these indicators in multivariate analyses but also to control and check for interactions with the survey mode when analyzing the data. Particularly when surveys are changing from face-to-face mode to web or mixed mode, due to adaption to the COVID-19 pandemic or simply due to cost-efficiency, data users should take the mode of conducting interviews into account. This is especially relevant when central variables measure subjective and sensitive assessments and are prone to social desirability bias. Otherwise, researchers can run the risk of confounding mode effects with substantive effects – for example, in terms of cross-national differences or change over time.

6. Acknowledgments

This project was funded by the Horizon 2020 Research and Innovation Programme under grant agreement 739511 for the project Generations and Gender Programme: Evaluate, Plan, Initiate and by the Federal Institute for Population Research (BiB) in Wiesbaden, Germany, to co-finance the German data collection.

We would like to thank the GGP pilot team (Tom Emery, Susana Cabaço, Peter Lugtig, Vera Toepoel, Martin Bujard, and Robert Naderi) for designing and conducting the pilot study, as well as Tobias Gummer, Karsten Hank, Sandra Krapf, and two anonymous reviewers for their helpful comments and suggestions, which assisted us greatly.

References

AAPOR (2016). The American Association for Public Opinion Research Survey Outcome Rate Calculator 4.1 [electronic resource]. https://www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Response-Rates-An-Overview.aspx.

Aquilino, W.S. (1991). Telephone versus face-to-face interviewing for household drug use surveys. *International Journal of the Addictions* 27(1): 71–91. doi:10.3109/10826089109063463.

Arránz-Becker, O. (2013). Effects of similarity of life goals, values, and personality on relationship satisfaction and stability: Findings from a two-wave panel study. *Personal Relationships* 20: 443–461. doi:10.1111/j.1475-6811.2012.01417.x.

Atkeson, L., Adams, A., and Alvarez, R. (2014). Nonresponse and mode effects in self- and interviewer-administered surveys. *Political Analysis* 22(3): 304–320. doi:10.1093/pan/mpt049.

Bethlehem, J. and Biffignandi, S. (2012). *Handbook of web surveys.* Hoboken: Wiley. doi:10.1002/9781118121757.

Bethlehem, J. and Stoop, I. (2007). Online panels – A paradigm theft? In: Trotman, M. et al. (eds.). *The challenges of a changing world. Proceedings of the Fifth ASC International Conference*. University of Southampton, 12–14 September 2007. Berkeley: Association for Survey Computing: 113–132.

Beullens, K., Loosveldt, G., Vandenplas, C., and Stoop, I. (2018). Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field*. doi:10.13094/SMIF-2018-00003.

Bradbury, T.N., Fincham, F.D., and Beach, S.R.H. (2000). Research on the nature and determinants of marital satisfaction: A decade in review. *Journal of Marriage and the Family* 62(4): 964–980. doi:10.1111/j.1741-3737.2000.00964.x.

Braekman, E., Charafeddine, R., Demarest, S., Drieskens, S., Berete, F., Gisle, L., Van der Heyden, J., and Van Hal, G. (2020). Comparing web-based versus face-to-face and paper-and-pencil questionnaire data collected through two Belgian health surveys. *International Journal of Public Health* 65: 5–16. doi:10.1007/s00038-019-01327-9.

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., Conrad, F., Wellings, K., Johnson, A.M., and Erens, B. (2016). Using the web to collect data on sensitive behaviours: A study looking at mode effects on the British National Survey of Sexual Attitudes and Lifestyles. *PLoS ONE* 11(2): 1–12. doi:10.1371/journal.pone.0147983.

Callegaro, M. (2008). Social desirability. In: Lavrakas, P.J. (ed.). *Encyclopedia of survey research methods*. Thousand Oaks: Sage: 825–826.

Chang, L. and Krosnick, J.A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly* 74(1): 154–167. doi:10.1093/poq/nfp090.

Christensen, A.I., Ekholm, O., Glümer, C., and Juel, K. (2013). Effect of survey mode on response patterns: Comparison of face-to-face and self-administered modes in health surveys. *European Journal of Public Health* 24(2): 327–332. doi:10.1093/eurpub/ckt067.

Couper, M.P. (2011). The future of modes of data collection. *Public Opinion Quarterly* 75(5): 889–908. doi:10.1093/poq/nfr046.

Couper, M.P., Antoun, C., and Mavletova, A. (2017). Mobile web surveys. In: Biemer, P.P., De Leeuw, E.D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L.E., Tucker, N.C., and West, B.T. (eds). *Total survey error in practice*. New Jersey: Wiley: 133–150. doi:10.1002/9781119041702.ch7.

Daikeler, J., Bošnjak, M., and Manfreda, K.L. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology* 8(3): 513–539. doi:10.1093/jssam/smz008.

Davis, R.E., Couper, M.P., Janz, N.K., Caldwell, C.H., and Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research* 25(1): 14–26. doi:10.1093/her/cyp046.

De Leeuw, E.D., Hox, J., and Dillman, D. (2008). *International handbook of survey methodology*. New York: Psychology Press.

DeMaio, T.J. (1984). Social desirability and survey measurement: A review. In: Turner, C.F. and Martin, E. (eds.). *Surveying subjective phenomena 2*. New York: Russel Sage Foundation: 257–282.

Emery, T., Cabaço, S., Lugtig, P., Toepoel, V., Lück, D., Naderi, R., Bujard, M., and Schumann, A. (2018). The Generations and Gender Programme: Evaluate, plan, initiate (Deliverable 2.1: GGP Technical Case and E-Needs). doi:10.31235/osf.io/439wc.

Fincham, F.D. and Beach, S.R.H. (2006). Relationship satisfaction. In: Vangelisti, A.L. and Perlman, D. (eds.). *The Cambridge handbook of personal relationships*. Cambridge: Cambridge University Press: 579–594. doi:10.1017/CBO9780511606632.032.

Fu, H., Darroch, J.E., Henshaw, S.K., and Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives* 30(3): 128–138. doi:10.2307/2991627.

Gauthier, A.H., Cabaço, S.L.F., and Emery, T. (2018). Generations and Gender Survey study profile. *Longitudinal and Life Course Studies* 9(4): 456–465. doi:10.14301/llcs.v9i4.500.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004). *Survey methodology*. New Jersey: Wiley.

Gummer, T., Höhne, J.K., Rettig, T., Roßmann, J., and Kummerow, M. (2023). Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. *Quality and Quantity*: 1–21. doi:10.1007/s11135-022-01601-8.

Gummer, T., Schmiedeberg, C., Bujard, M., Christmann, P., Hank, K., Kunz, T., Lück, D., and Neyer, F.J. (2020). The impact of COVID-19 on fieldwork efforts and planning in pairfam and FReDA-GGS. *Survey, Research, Methods* 14(2): 223–227. doi:10.18148/srm/2020.v14i2.7740.

Gustavson, K., Nilsen, W., Ørstavik, R., and Røysamb, E. (2013). Relationship quality, divorce, and well-being: Findings from a three-year longitudinal study. *The Journal of Positive Psychology* 9(2): 163–174. doi:10.1080/17439760.2013.858274.

Heerwegh, D. (2009). Mode differences between face-to-face and web-surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research* 21(1): 111–121. doi:10.1093/ijpor/edn054.

Heerwegh, D. and Loosfeldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population. Differences in response quality. *Public Opinion Quarterly* 72(5): 836–846. doi:10.1093/poq/nfn045.

Holbrook, A.L., Green, M.C., and Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly* 67(1): 79–125. doi:10.1086/346010.

Huß, B. and Pollmann-Schult, M. (2020). Relationship satisfaction across the transition to parenthood: The impact of conflict behavior. *Journal of Family Issues* 41(3): 383–411. doi:10.1177/0192513X19876084.

Kalmijn, M. (2021). Are national family surveys biased toward the happy family? A multi-actor analysis of selective survey nonresponse. *Sociological Methods and Research* online first: 1–26. doi:10.1177/0049124120986208.

Kantar Public (2018). Beziehungen und Familienleben in Deutschland (pairfam). Methodenbericht Welle 10(2017/2018). Kantar Public.

Karney, B.R. and Bradbury, T.N. (1995). The longitudinal course of marital quality and stability. *Psychological Bulletin* 118(1): 3–34. doi:10.1037/0033-2909.118.1.3.

Karney, B.R. and Bradbury, T.N. (2020). Research on marital satisfaction and stability in the 2010s: Challenging conventional wisdom. *Journal of Marriage and Family* 82(1): 100–116. doi:10.1111/jomf.12635.

Kelly, C.A., Soler-Hampejsek, E., Mensch, B.S., and Hewett, P.C. (2013). Social desirability bias in sexual behavior reporting: Evidence from an interview mode experiment in rural Malawi. *International Perspectives on Sexual and Reproductive Health* 39(1): 14–21. doi:10.1363/3901413.

Kluwer, E.S. and Johnson, M.D. (2007). Conflict frequency and relationship quality across the transition to parenthood. *Journal of Marriage and Family* 69(5): 1089–1106. doi:10.1111/j.1741-3737.2007.00434.x.

Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys. The effects of mode and question sensitivity. *Public Opinion Quarterly* 72(5): 847–865. doi:10.1093/poq/nfn063.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality Quantity* 47: 2025–2047. doi:10.1007/s11135-011-9640-9.

Lee, R.M. and Renzetti, C.M. (1990). The problems of researching sensitive topics: An overview and introduction. *American Behavioral Scientist* 33(5): 510–528. doi:10.1177/0002764290033005002.

Lewis, R.A. and Spanier, G.B. (1979). Theorizing about the quality and stability of marriage. In: Burr, W.R., Hill, R.F., Nye, I., and Reiss, I.L. (eds.). *Contemporary theories about the family*. New York: Free Press: 268–293.

Liu, M. (2017). Data collection mode differences between national face-to-face and web surveys on gender inequality and discrimination questions. *Women's Studies International Forum* 60: 11–16. doi:10.1016/j.wsif.2016.11.007.

Liu, M. and Stainback, K. (2013). Interviewer gender effects on survey responses to marriage-related questions. *Public Opinion Quarterly* 77(2): 606–618. doi:10.1093/poq/nft019.

Liu, M. and Wang, Y. (2016). Comparison of face-to- face and web surveys on the topic of homosexual rights. *Journal of Homosexuality* 63(6): 838–854. doi:10.1080/00918369.2015.1112587.

Lugtig, P., Toepoel, V., Emery, T., Cabaço, S.L.F., Bujard, M., Naderi, R., Schumann, A., and 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. SocArXiv. doi:10.31235/osf.io/mu8jy.

Medway, R. (2012). Beyond response rates: The effect of prepaid incentives on measurement error [PhD thesis]. Maryland: University of Maryland.

Rijken, A.J. and Liefbroer, A.C. (2009). The influence of partner relationship quality on fertility. *European Journal of Population* 25: 27–44. doi:10.1007/s10680-008-9156-8.

Robertson, R.E., Tran, F.W., Lewark, L.N., and Epstein, R. (2018). Estimates of non-heterosexual prevalence: The roles of anonymity and privacy in survey methodology. *Archive of Sexual Behavior* 47: 1069–1084. doi:10.1007/s10508-017-1044-z.

Schmid, L., Wörn, J., Hank, K., Sawatzki, B., and Walper, S. (2021). Changes in employment and relationship satisfaction in times of the COVID-19 pandemic: Evidence from the German family Panel. *European Societies* 23(sup1: European Societies in the Time of the Coronavirus Crisis): S743–S758. doi:10.1080/14616696.2020.1836385.

Schonlau, M. and Couper, M.P. (2017). Options for conducting web surveys. *Statistical Science* 32(2): 279–292. doi:10.1214/16-STS597.

Schonlau, M., van Soest, A., and Kapteyn, A. (2007). Are 'Webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods* 1(3): 155–163. doi:10.2139/ssrn.1006108.

Schonlau, M., van Soest, A., Kapteyn, A., and Couper, M.P. (2009): Selection bias in web surveys and the use of propensity scores. *Sociological Methods and Research* 27(3): 291–318. doi:10.1177/0049124108327128.

Schröder, J. and Schmiedeberg, C. (2020). Effects of partner presence during the interview on survey responses: The example of questions concerning the division of household labor. *Sociological Methods and Research* online first: 1–23. doi:10.1177/0049124120914938.

Tourangeau, R. (2017). Presidential address. Paradoxes of nonresponse. *Public Opinion Quarterly* 81(3): 803–814. doi:10.1093/poq/nfx031.

Tourangeau, R. and Smith, T.W. (1996). Asking sensitive questions. The impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 60(1): 275–304. doi:10.1086/297751.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin* 133(5): 859–883. doi:10.1037/0033-2909.133.5.859.

Van Damme, M. and Dykstra, P. (2018). Spousal resources and relationship quality in eight European countries. *Community, Work and Family* 21(5): 541–563. doi:10.1080/13668803.2018.1526776.

Vehovar, V., Batagelj, Z., Manfreda, K.L., and Zaletel, M. (2002). Nonresponse in web surveys. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds.). *Survey Nonresponse*. New York: Wiley: 229–242.

Weinreb, A., Sana, M., and Stecklov, G. (2018). Strangers in the field: A methodological experiment on interviewer–respondent familiarity. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 137–138(1): 94–119. doi:10.1177/0759106318761562.

Wiik, K.A., Keizer, R., and Lappegard, T. (2012). Relationship quality in marital and cohabiting unions across Europe. *Journal of Marriage and Family* 74(3): 389–398. doi:10.1111/j.1741-3737.2012.00967.x.

Wolf, C., Christmann, P., Gummer, T., Schnaudt, C., and Verhoeven, S. (2021). Conducting general social surveys as self-administered mixed-mode surveys. *Public Opinion Quarterly* 85(2): 623–648. doi:10.1093/poq/nfab039.

Appendices

Table A-1: Distributions and sources of reference information from official statistics used for post-stratification weighting

Indicator	Source	Year of conduction	Age of population (in years)	Categories	Distribution (in percent)
Sex	Micro census	2016	18–49	(1) Male	(1) 51.63
				(2) Female	(2) 48.37
Age	Census	2011	18–49	(1) 18–29	(1) 32.95
				(2) 30–39	(2) 28.02
				(3) 40–49	(3) 39.02
Education	Census	2011	18–49	(1) Not (yet) graduated/low education (Hauptschulabschluss)	(1) 38.08
				(2) Middle school education (Realschulabschluss)	(2) 30.55
				(3) High school education (Fachhochschulabschluss/Abitur)	(3) 31.38
Nationality	Census	2011	18–49	(1) German	(1) 90.71
				(2) Non-German	(2) 9.29
Regional setting	Registry office	2018	18–49	(1) Rural area	(1) 50.00
				(2) Urban area	(2) 50.00

Source: Micro census 2016, census 2011, and register data; authors' own calculations.

TableA-2: Original wordings of questions and answers for items measuring relationship quality in the GGP pilot study 2018

Item	Question text	Answer categories
a220	Even people who get along well with their partners sometimes wonder whether their marriage or partnership will work. Over the past 12 months, have you thought about breaking up your relationship?	Yes No
a217	How satisfied are you with your relationship with your partner/spouse?	On a scale from 0 to 10, where 0 means "not at all satisfied," 10 means "completely satisfied." and 5 means "about average," what number best represents your satisfaction? 0–10
a312	How satisfied are you with the division of household tasks between you and your partner/spouse?	
a314	How satisfied are you with the way child care tasks are divided between you and your partner/spouse?	
a218	In the last 12 months, how often did you have disagreements with your partner about:	Never Seldom Sometimes Frequently Very frequently
a218a	household chores?	
a218b	money?	
a218c	use of leisure time?	
a218d	relations with friends?	
a218e	relations with parents?	
a218f	having children?	
a218g	child-raising issues?	
a219	Couples deal with serious disagreements in very different ways. If you have a serious disagreement with your partner, how often do you:	Never Seldom Sometimes Frequently Very Frequently
a219a	avoid discussion by giving in?	
a219b	discuss your disagreement calmly?	
a219c	argue heatedly or shout?	
a219d	refuse to talk about it?	

Source: GGP pilot study 2018.

Table A-3: Post-stratification weighted means of indicators or percentage of confirmative answers by mode of data collection

	F2F	Web	Mode difference	F2F	Web
	Mean or percentage		Δ	n	
Subjective instability [a]					
Thought about breaking up (in %)	6.28 (3.53–10.93)	14.49 (9.55–21.40)	**8.21**	144	175
Satisfaction [b]					
General relationship	9.15 (8.96–9.35)	8.71 (8.48–8.94)	**–0.44**	144	185
Household tasks	8.47 (8.16–8.78)	7.88 (7.50–8.25)	**–0.59**	126	160
Child care tasks	8.63 (8.24–9.02)	8.24 (7.78–8.69)	**–0.39**	70	79
Conflict frequency [c]					
Household chores	2.30 (2.12–2.49)	2.42 (2.26–2.57)	**0.12**	145	188
Money	1.68 (1.52–1.83)	2.05 (1.85–2.26)	**0.37**	145	188
Leisure time	2.25 (2.06–2.44)	2.38 (2.22–2.54)	**0.13**	145	188
Relations with friends	1.52 (1.39–1.65)	1.71 (1.55–1.86)	**0.19**	144	188
Relations with parents	1.68 (1.52–1.83)	1.84 (1.69–2.00)	**0.16**	145	188
Having children	1.30 (1.16–1.44)	1.23 (1.14–1.33)	**0.07**	144	185
Child-raising issues	1.81 (1.63–2.00)	1.89 (1.70–2.08)	**0.08**	141	174
Conflict style [d]					
Avoid discussion by giving in	2.71 (2.53–2.91)	2.67 (2.50–2.84)	**–0.04**	131	177
Discuss conflicts calmly	4.03 (3.85–4.20)	3.95 (3.78–4.12)	**–0.08**	135	175
Argue heatedly or get loud	1.91 (1.77–2.05)	2.06 (1.88–2.23)	**0.15**	135	179
Refuse to talk	1.71 (1.55–1.87)	1.83 (1.67–1.99)	**0.12**	135	177

[a] Reference category is "not thought about breakup."
[b] From 0 (very dissatisfied) to 10 (very satisfied).
[c] From 1 (never) to 5 (very frequently).
[d] From 1 (never) to 5 (very frequently).
Notes: F2F = face-to-face; Δ = mode difference (in bold); 95% confidence intervals in parentheses.
Source: GGP pilot study 2018; authors' own calculations.

Table A-4: Effect of web mode on separate indicators of relationship quality under block-wise adjustment of demographic and family-related confounders

	Model 1: Baseline		Model 2: Demographic controls		Model 3: Demographic and family-related controls	
Subjective Instability	Linear probability models					
	Coef.	SE	Coef.	SE	Coef.	SE
Mode *(Ref: Face-to-face)*						
Web	0.082 (0.009–0.154)	0.037	0.087 (0.014–0.161)	0.037	0.094 (0.020–0.167)	0.037
n	326		319		317	
df	1		6		9	
R^2	0.01		0.04		0.06	
Adjusted R^2	0.01		0.02		0.04	
Subjective Instability	Logistic regression models					
	AME	SE	AME	SE	AME	SE
Mode *(Ref: Face-to-face)*						
Web	0.082 (0.099–0.153)	0.037	0.088 (0.016–0.161)	0.037	0.095 (0.022–0.168)	0.037
n	326		319		317	
df	1		6		9	
AIC	257.26		256.99		254.38	
BIC	264.84		283.35		291.98	
Satisfaction – relationship	Linear regression models					
	Coef.	SE	Coef.	SE	Coef.	SE
Mode *(Ref: Face-to-face)*						
Web	–0.484 (–0.769 – –0.198)	0.143	–0.508 (–0.789 – –0.227)	0.143	–0.533 (–0.817 – –0.250)	0.144
n	335		329		327	
df	1		6		9	
R^2	.03		.11		.12	
Adjusted R^2	.03		.09		.10	
Satisfaction – household	Linear regression models					
	Coef.	SE	Coef.	SE	Coef.	SE
Mode *(Ref: Face-to-face)*						
Web	–0.559 (–0.980 – –0.138)	0.214	–0.578 (–0.998 – –0.158)	0.213	–0.617 (–1.040 – –0.194)	0.215
n	293		286		285	
df	1		6		8	
R^2	.02		.09		.10	
Adjusted R^2	.02		.07		.07	

Table A-4: (Continued)

	Model 1: Baseline		Model 2: Demographic controls		Model 3: Demographic and family-related controls	
Satisfaction child care	Linear regression models					
	Coef.	SE	Coef.	SE	Coef.	SE
Mode *(Ref: Face-to-face)*						
Web	–0.546 (–1.083 – –0.009)	0.272	–0.526 (–1.060 – –0.007)	0.270	–0.508 (–1.042 – 0.027)	0.270
n	151		149		149	
df	1		6		7	
R^2	.03		.11		.11	
Adjusted R^2	.02		.07		.07	
Conflict frequency	Linear regression models					
	Coef.	SE	Coef.	SE	Coef.	SE
Mode *(Ref: Face-to-face)*						
web	1.110 (0.276–1.944)	0.424	1.121 (0.273–1.968)	0.431	1.268 (0.434–2.102)	0.424
n	320		313		311	
df	1		6		9	
R^2	.02		.04		.10	
Adjusted R^2	.02		.02		.008	
Conflict style	Linear regression models					
	Coef.	SE	Coef.	SE	Coef.	SE
Mode *(Ref: Face-to-face)*						
web	0.634 (0.111–1.156)	0.266	0.676 (0.148–1.204)	0.268	0.717 (0.185–1.249)	0.270
n	309		304		302	
df	1		6		9	
R^2	.02		.04		.05	
Adjusted R^2	.02		.02		.02	

Notes: Model 1: baseline model; Model 2: under-adjustment of demographic variables (sex, nationality, age, education, and regional setting); Model 3: under-adjustment of family-related variables (relationship status, coresident children under 6). Ref = reference category; n = observations; df = degrees of freedom; AME = average marginal effects; Coef. = coefficients; SE = standard error; 95% confidence intervals in parentheses.
Source: GGP pilot study 2018; authors' own calculations.