Survey Research Methods (2025)
Vol. 19, No. 2, pp. 139-152
https://doi.org/10.18148/srm/2025.v19i2.8348
European Survey Research Association

Surely Shorter is Better? A Questionnaire Length Experiment in a Self-completion Survey

Tim Hanson[1] · Eva Aizpurua[2] · Rory Fitzgerald[1] · Marta Vukovic[3]
[1]City St George's, University of London
[2]National Centre for Social Research
[3]University of Vienna

As surveys increasingly transition to self-completion approaches, understanding the impact of design decisions on survey outcomes becomes paramount. This includes assessing the effects of different questionnaire lengths on survey quality and considering what length of questionnaire can reasonably be fielded in a self-completion environment. This paper presents findings from a self-completion experiment conducted in Austria in 2021, which compared the full European Social Survey Round 10 questionnaire (estimated 50-minute completion time) with a shorter version (anticipated 35-minute completion time). The analysis includes a comparison between the longer and shorter versions based on response rates, sample composition, and data quality indicators. Except for response rates, which were significantly but only slightly higher in the shorter condition, results for both versions were generally comparable. Fielding a shorter questionnaire did not produce clear benefits in terms of sample composition or data quality. These results suggest that a questionnaire that would traditionally be regarded as lengthy for self-completion can yield acceptable outcomes.

Keywords: interview length; self-administered survey; european social survey; response rates; data quality

1 Introduction

In many countries, face-to-face surveys have seen escalating costs (Wolf, Christmann, Gummer, Schnaudt & Verhoeven, 2021) and consistent declines in response rates (Beullens, Loosveldt, Vandenplas & Stoop, 2018; Luiten, Hox & de Leeuw, 2020). At the same time, concerns have grown about the impact of negative interviewer effects, especially where these differ between nations in cross-national studies (Loosveldt & Beullens, 2017; Loosveldt & Wuyts, 2020). While these challenges are not new, the COVID-19 pandemic introduced additional difficulties for face-to-face data collection, accelerating the need to develop high-quality alternatives, including self-completion modes (Olson et al., 2020). However, self-completion surveys come with their own challenges, such as lower response rates (Daikeler, Bosnjak & Manfreda, 2020), difficulties with selection of respondents within households (Smyth, Olson & Stange, 2019), concerns about administering lengthy questionnaires (Revilla & Höhne, 2020), lower respondent attention (Anduiza & Galais, 2017), and other concerns including coverage issues in web surveys, exclusion of low-literacy groups, and questionnaire administration limitations for paper self-completion instruments.

This paper examines the feasibility of administering longer surveys using self-completion instruments by testing two survey lengths that are traditionally fielded as face-to-face surveys. Our analysis is based on a self-completion (web followed by paper) experiment conducted in Austria in 2021, using the European Social Survey (ESS) Round 10 questionnaire. We compare response rates, sample composition, and data quality indicators between approximate

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2025.v19i2.8348) contains supplementary information.

Corresponding author: Tim Hanson, City St George's, University of London, London, England
(Email: tim.hanson@citystgeorges.ac.uk)

questionnaire lengths of 50 and 35 min. This research builds on an earlier experiment conducted in three countries (Austria, Hungary, Serbia) that tested a shorter version of the ESS questionnaire (≅ 20 min) with positive outcomes (Fitzgerald, 2021).

2 Background

2.1 The Relevance of Survey Length

Peytchev et al. (2020) observe that the impact of survey length on survey participation has received surprisingly little attention, perhaps being seen more as a static attribute than an alterable design feature. While survey length can be shortened by reducing the number of questions, such reduction comes with trade-offs, reducing the scope of the survey and providing less data to users. Therefore, the impact of longer questionnaires on survey outcomes warrants thorough examination to inform decisions about the nature and scale of potential reductions in content.

The length of a questionnaire can influence multiple dimensions of response: respondents' willingness to start the survey, their likelihood of completing it, and the quality of the data they provide. Singer's (2011) benefit-cost theory of survey participation argues that individuals choose to participate when they perceive the benefits of doing so outweigh the costs. As the length of a survey increases, the perceived 'cost' of participation may be higher, and consideration may be needed as to whether the benefits (financial or otherwise) are sufficient, or can be made sufficient, to counterbalance this.

Leverage salience theory, introduced by Groves, Singer, and Corning (2000), describes how various survey design attributes have different 'leverages' on the decision to cooperate for different people. Extending this to survey length, this theory suggests that a longer survey may deter participation for certain sample members, even if a difference in length may be less important for other people.

Any discussion on survey length links with the extensive literature on response burden. Yan and Williams (2022) note that 'burden' is used as a broad term to refer to both actual (objective) and perceived (subjective) burden. Bradburn (1977) identified four factors that influence response burden: (1) interview length, (2) the effort required by respondents, (3) the frequency of interviewing, and (4) the amount of stress experienced by respondents. Gummer and Roßmann (2015) observe that interview duration is a direct measure of response burden for respondents, with longer surveys typically imposing a higher burden on respondents.

There may also be an interaction between survey mode and questionnaire length. A longer questionnaire might be seen as more challenging in a self-completion survey compared to a face-to-face interview, where the interviewer has sometimes been considered more capable of motivating target respondents to complete the survey (Luijkx et al., 2021).

2.2 Impact of Survey Length on Response Rates

A low response rate may increase the risk of non-response bias, though the extent to which this applies can vary from survey to survey (Groves et al., 2008). A survey's response rate is determined both by the proportion of (eligible) respondents who start the survey and the proportion of those who, having started, complete it.

The announced survey duration might influence the decision of potential respondents to start the survey. If the survey is perceived as being excessively long, some may deem it too burdensome, and elect not to start it. Past research has shown that the longer the stated duration is for a web survey, the fewer respondents start it. This includes Galesic and Bosnjak (2009), who compared the impact of announced lengths of 10, 20, and 30 min for an opt-in web survey in Croatia. Their findings revealed that 75% started the 10-minute survey, compared with 65% for the 20-minute survey and 62% for the 30-minute survey. Similarly, Yan, Conrad, Tourangeau and Couper (2011) found that fewer individuals started a web survey when presented with a longer estimated duration (25 or 40 min) than a shorter one (5 or 10 min).

Other studies have examined the relationship between survey length and break-off rates (the proportion of individuals who started the survey but did not finish it). Yan et al. (2011, p. 141) found higher break-off rates for a 25-minute survey compared with a 16-minute one, noting this was to be anticipated as "each new question provides an opportunity to break off and there are more opportunities in the long than short questionnaire."

Mavletova and Couper (2015) conducted a meta-analysis of 39 independent samples and found lower breakoff rates in shorter surveys when respondents were using mobile devices to complete web surveys. This group is particularly important given the growth in use of mobile devices for survey completion (Gummer, Höhne, Rettig, Roßmann and Kummerow, 2023) and the greater likelihood of those using these devices to break-off midway through surveys (Emery et al., 2023).

Further evidence is provided by Marcus, Bosnjak, Lindner, Pilischenko and Schütz (2007), who assessed the overall response rate of a web survey comparing two lengths—one version containing 91 items and one encompassing 359 items. The response rate for the longer version was 19%, significantly lower than the 31% achieved by the shorter version.

While a significant body of evidence points to the relationship between questionnaire length and response rates, the results across self-completion studies are not always consistent. Rolstad, Adler and Rydén (2011) carried out a meta-analysis of 25 paper surveys comparing different questionnaire lengths based on the number of pages. Of the studies examined, only six reported a significant reduction in response rates when longer questionnaires were administered. They conclude that other factors (e.g., survey content) may be as important as the length of the questionnaire. Yan et al. (2022, p. 951), referencing this study, observed that "... it is not clear whether or not survey length actually does lead to response burden and lower response rates and what the survey field (and regulators) should do with regard to survey length".

2.3 Impact of Survey Length on Data Quality

The length of a questionnaire may also impact the quality of the data provided by respondents. Krosnick (1991) introduced the term 'satisficing' to describe situations where respondents provide a satisfactory rather than an optimal answer, often due to the increased cognitive effort required for the latter. He identifies three key factors that contribute to satisficing: (1) task difficulty, (2) respondent ability, and (3) respondent motivation. It might be hypothesised that questionnaire length directly affects both task difficulty and respondent motivation. A longer questionnaire may impose a greater cumulative burden to the respondent's task and result in increased satisficing behaviours, particularly towards the end of the questionnaire. Equally, respondents may be less motivated to put optimal effort in completing a longer survey.

Peytchev and Peytcheva (2017) note that researchers have long suspected a positive association between survey length and measurement error, citing two types of evidence to support this claim. The first involves respondents providing answers that will deliberately allow them to skip additional questions (Daikeler, Bach, Silber & Eckman, 2022). The second is non-random measurement error caused by lower quality answers towards the end of surveys. Examples of the latter include Galesic et al. (2009) who found that answers to questions located later in the questionnaire were faster, shorter, and more uniform than those positioned near the beginning. Sahlqvist et al. (2011) compared a 15- and a 24-page paper questionnaire and found that item non-response was significantly higher in the longer questionnaire compared with the shorter one (10% versus 6%). However, it remains unclear if this is partly related to the omission of certain types of question from the shorter questionnaire, where non-response may have been higher.

These examples offer some tentative evidence of the relationship between survey length and data quality. However, evidence in this area remains limited, underscoring the need for further research. Our study contributes to this body of research by comparing data quality across multiple indicators for a subset of questions asked in both questionnaire length versions. Such a comparison will help establish whether any decline in response quality over the course of a survey is heightened as the overall length increases.

2.4 What Might be the Different Impacts for Web and Paper Questionnaires?

In the above sections, we have noted several studies examining the impact of questionnaire length on response rates and data quality in self-completion surveys. Such surveys can either be administered via the web, on paper, or a combination of both, and the effects may differ depending on which self-completion mode(s) are used. Most of the referenced experiments based on web surveys show a decline in response rates as survey length increases. Rolstad's et al. (2011) meta-analysis stands out as an exception to this general finding, with only a quarter of the studies showing significantly lower response rates for longer questionnaires. However, this analysis was based on paper questionnaires, suggesting that the implications of survey length could manifest differently between paper and web modes.

There are clear differences in the way respondents interact with surveys depending on the mode. For example, the length of the survey will be more immediately apparent to those receiving a paper version as they can see the size of the questionnaire. They can also see the questionnaire routing and how their answers lead to subsequent questions. There may be differences between how long different respondents can engage with each mode and how likely they are to take breaks while completing the questionnaire. For mixed-mode surveys that allow both web and paper responses, there are also differences in who responds using each mode. For example, younger and more educated individuals are more likely to respond on the web whereas older people are more likely to take part on paper (Kotnarowski, 2023). If survey length has differential impacts between web and paper questionnaires, this can impact on the achieved sample composition. To further assess these interactions, it is important to compare the impact of different questionnaire lengths between web and paper instruments, as we do in our experiment.

2.5 Why Might Longer Questionnaires Be Needed?

Revilla et al. (2020) asked panellists from two online panels in Germany about their opinions on the ideal and maximum lengths of surveys. Their results suggest that the ideal length for an online survey is between 10 and 15 min, with the maximum length ranging from 20–28 min. The authors conclude that their study can provide valuable information for researchers and practitioners in deciding the length of online surveys, noting that surveys exceeding 30 min are often perceived as overly lengthy.

The general advice to limit the length of questionnaires seems logical. A questionnaire should only include the questions needed to meet the study's objectives, avoiding unnecessary or lower-priority content that could increase respondent burden. However, the challenge can be greater for existing, lengthy questionnaires that are being transitioned to new data collection modes. For example, the European Social Survey (ESS), which was conducted face-to-face for its first nine rounds (2002–2018), typically takes about 60 min to complete. Such a duration is not unusual for face-to-face surveys. If questionnaires longer than 30 min prove problematic in self-completion environments, the ESS, along with other similar surveys, would require substantial modifications before transitioning to a self-administered approach.

Trimming a questionnaire by potentially half means certain topics may need to be removed, thereby diminishing the value of the data set for both users and sponsors. There are considerable costs and efforts required to field a high-quality cross-national survey and justifying this investment may become harder when the amount of data collected is reduced. Having additional but shorter surveys can also be operationally challenging in a cross-national setting.

Efforts have been made to limit questionnaire length with minimal content loss by adopting matrix or modularised designs. A matrix approach involves a planned missing data design (Enders, 2010) that seeks to reduce respondent burden by distributing questionnaire items across different versions, with each respondent receiving only a subset of the total items. Luijkx et al. (2021) tested a matrix design in the 2017 European Values Study, where the questionnaire was split into shorter versions, each containing a subset of questions from the full questionnaire. Respondents were then randomly assigned to different versions, ensuring that all questions were answered by a sufficiently large number of respondents across the entire sample. However, they found similar response rates for both the full and matrix versions.

A modularised design involves splitting a survey into multiple parts which are completed by respondents in separate shorter 'chunks', rather than as a single exercise. Peytchev et al. (2020) tested a modularised design, in which a 30-minute questionnaire was split into two 15-minute parts, with respondents asked to complete each part separately. They found that the single 30-minute version yielded a higher response rate than the combined rate of the two 15-minute parts.

The matrix and modularised designs highlighted above are only two examples, and it might be possible that in other contexts such designs could deliver improved outcomes compared with a single longer questionnaire version. However, even if such designs were to improve response rates and other metrics, they would also introduce challenges. For example, a matrix design results in a more complex data structure, with data available at different levels for different parts of the questionnaire. This also limits opportunities for analysis of relationships between topics covered in different versions. The ESS has always sought to produce accessible data sets that can be easily used by all levels of users, and designs that add complexity may reduce this accessibility. Furthermore, administering multiple questionnaire versions in a cross-national survey covering around 30 countries, which also includes a paper questionnaire, presents clear practical challenges.

While we acknowledge the potential benefits of matrix and modularised designs, they are not currently considered as an option for the ESS. Therefore, our focus for this paper is on the outcomes associated with fielding a complete version of the ESS questionnaire in a self-completion context versus a shorter version.

Given the drawbacks of limiting questionnaire length, there is a need for a robust evidence base to inform sponsors and practitioners about the repercussions of different survey lengths. Ultimately, this consideration may go beyond a comparison of which survey length delivers the best response rate or sample composition. It may be the case, for example, that a shorter questionnaire does achieve a better response rate than a longer version. However, this improvement might come at the cost of a reduction in the data delivered from the survey, which can impact on its overall utility. It may therefore need to be considered whether a small decline in response rate is an acceptable trade-off for the additional data that is provided to data users and funders. This may be particularly the case if there are no differences in sample composition or data quality between the different length versions.

There are encouraging findings from cross-national studies probing the feasibility of transitioning long surveys from interviewer-administered to self-administered approaches. The 2017 European Values Study (approx. length 60 min) achieved responses rates ranging between 35 and 45% in Germany, Switzerland, Denmark, and Iceland when testing its full survey questionnaire in self-completion modes (Luijkx et al., 2021). A three-country experiment in the Gender and Generations Survey (approx. length 40 min) found that

a push-to-web design was equally or more successful than a face-to-face design in two of the three countries (Croatia and Germany, respective push-to-web response rates of 48% and 26%) (Lugtig et al., 2022). However, the push-to-web design was less successful in Portugal (13% response rate). In Round 10 of the European Social Survey, due to restrictions on face-to-face fieldwork from the COVID pandemic, nine countries adopted a self-completion (web and paper) approach, based on the near-full (60-minute) questionnaire. In 7 of the 9 countries, response rates ranging between 30 and 39% were achieved (O'Muircheartaigh, Hanson & Fitzgerald, 2023). These examples provide tentative evidence that reasonable response rates can be achieved for long questionnaires in a self-completion context. However, further research is needed regarding how these figures compare with what can be achieved with shorter questionnaires, and to assess impacts beyond headline response rates.

3 Current Study

The current study adds to the existing evidence on the impact of survey length for self-completion surveys in three important respects. First, our experiment is based on a long survey, averaging nearly one hour for the face-to-face ESS. With a few exceptions, most of the experiments in the literature are based on shorter lengths (e.g., 5–30 min). There is a marked difference between comparing a 5-minute survey with a 10-minute survey, versus comparing two lengthier versions (in this case, a 35- and 50-minute questionnaire). It was a deliberate decision to compare two versions, both of which might be seen as long in a self-completion survey context, as reducing some survey questionnaires to under 30 min, for example, may be highly problematic.

Second, this experiment is based on a combination of web and paper modes. Many of the examples from the literature are based on either web or paper surveys, but rarely both. This experiment allows us to assess the impact on overall response rates as well as of the share of web and paper completion. This is important as it is not currently feasible in most European countries to deliver robust population level data from a web-only approach.

Third, in common with some of the above studies, we extend our analysis beyond a comparison of response rates between the different length versions to examine the impact on sample composition and data quality. This is important as the evidence suggests that response rates alone are a limited indicator of survey quality (Groves, et al., 2008). It may be the case, for example, that a longer questionnaire delivers a lower response rate, yet equivalent sample and data quality compared to a shorter version. Conversely, there may be no difference in response rates based on length, but data quality may be lower with a longer questionnaire.

3.1 Research Questions

This experiment was designed to answer the following research questions and test their accompanying hypotheses:

RQ_1 How does survey length influence response rates?

H_1 Response rates will be lower for the lengthier questionnaire. This is based on evidence from most of the studies cited in the literature review, which found lower response rates as questionnaire length increased. We also expect a higher breakoff rate for the longer questionnaire.

RQ_2 Are there differences in sample composition depending on survey length?

H_2 There is less evidence from previous studies regarding the impact of different survey lengths on sample composition. However, certain subgroups are often underrepresented in surveys, including younger people, those with lower levels of education, those on lower incomes, and those born outside of the country. We hypothesise that this underrepresentation may be compounded by the longer questionnaire length. Overall, we expect the sample composition to be closer to the population for the shorter condition.

RQ_3 Are there differences in data quality based on survey length?

H_3 We expect respondent burden to increase as questionnaire length increases, resulting in greater satisficing behaviours for the longer questionnaire. In particular, we expect item nonresponse and non-differentiation to be higher and concurrent validity and internal consistency to be lower in the lengthier questionnaire.

4 Methodology

4.1 Participants

The target population for this study comprised individuals aged 18 or over residing in private households in Austria. A two-stage approach was followed to sample target respondents. First, a random sample of 4000 households was drawn from the Austrian Postal Service (data.door). Since all contact with addresses was conducted via mail and no in-person visits were required, an unclustered sample was drawn.

Second, within sampled households, one adult aged 18 or over was (quasi)randomly selected. This selection was

done through instructions in the invitation letter and on the landing page, specifying that the survey should only be completed by the adult in the household with the next birthday. Check questions were included at the start of the questionnaire to try and ensure that the correct person was completing the survey.

4.2 Experimental Design and Procedure

Each sampled household was randomly assigned to receive either a 'shorter' (estimated length 35 min), or 'longer' (estimated length 50 min) version of the questionnaire. The expected survey length (35 or 50 min) was announced in the invitation letter. Actual mean completion times were about 10 min longer for both length versions. The questionnaire length experiment was crossed with an experiment on the conditional incentive amount. Within the shorter and longer conditions, respondents were randomly assigned to receive a €10 voucher following completion of the survey, or a €25 voucher. All groups received a €5 prepaid cash incentive with the first mailing. No significant interactions were found between length and incentive conditions. Therefore, for this paper, we combine the two shorter conditions and compare the results with the two longer conditions. This gives a gross sample size of 2000 cases per condition.

The lengthier version covered the full ESS Round 10 questionnaire, comprising 274 questions. The shorter version excluded two modules on democracy and digital social contacts. These two modules comprised 92 questions, thereby reducing the total question count for the shorter questionnaire to 182. These question counts refer to the total number of questions without taking into account routing. Consequently, the number of questions per respondent could be lower than this. Table A1 in Appendix 1 provides a summary of the questionnaire content and how this differed between the length conditions. Copies of the two paper questionnaire versions, as fielded in German language, are included in Annexe A. The lengthier questionnaire was 36 pages in the paper version while the shorter questionnaire was 28 pages.

Data collection was carried out between 13th April and 7th June 2021. After receiving approval from the Institutional Review Board, survey invitations were mailed on 13th April 2021. A reminder followed one week later on 20th April 2021, and a second reminder accompanied by a paper questionnaire was sent to all non-respondents a further two weeks later on 5th May 2021. A final postcard reminder was sent to those who had not responded on 18th May 2021. The web-based survey closed on 1st June, while the deadline for the return of paper questionnaires was 7th June. Data collection took place after COVID lockdowns in Austria had ended but while some restrictions were still in place (e.g., mandatory FFP2-mask wearing).

All letters included a simple URL (www.lebeninoesterreich.org) that led to a landing page which had a link to the Qualtrics web survey platform. To access the survey, participants were asked to enter a unique access code that was provided in the invitation and reminders. To encourage web responses, the first two mailings did not include a paper questionnaire, but a sentence was included highlighting that it would be sent in a later mailing. The questionnaire was provided in the German language.

We acknowledge that there were elements of the design used for this study may not reflect what was done for other studies that have compared questionnaire length between conditions. First, all sampled addresses were sent an unconditional monetary incentive (a €5 note). Combined with the conditional incentive (either €10 or €25), this may be a larger incentive than other studies in Europe have offered. Second, the invitation and reminder letters were jointly branded from the Institute for Advanced Studies (IHS), an independent research institute in Vienna, and the Institute for Empirical Social Research (IFES) in Austria. The letters were signed by Austria's National Coordinator for the European Social Survey, based at IHS. This branding may have given the survey greater academic credibility than may be an option for other studies. Third, the ESS questionnaire includes relatively little routing or complicated design elements, which makes it feasible to offer a paper questionnaire without needing excessive pages (as noted above, the 50-minute version equated to 36 pages). We also worked with Don Dillman as a consultant on the project and drew on his advice regarding the design of the questionnaire and materials (Dillman, Smyth and Christian, 2014). This included making a link between the survey topics mentioned in the letters and those appearing at the start of the questionnaire, varying the wording between letters regarding motivation to participate, and varying the size and style of envelopes for different mailings. Images of the survey materials, in German language, are included in Annexe B.

5 Measures

5.1 Response Rates (RQ1)

Response rates were calculated by dividing the number of valid responses by the number of valid addresses. Valid responses comprised fully and partially completed questionnaires with responses to at least one of the final 33 questions in the survey. The response rate calculation was based on the AAPOR RR1 formula, where the denominator for

the response rate represents the maximum number of potentially eligible cases, inclusive of those with unknown eligibility. Cases known to be ineligible, where invitation letters were returned as undeliverable, were excluded from the denominator.

5.2 Sample Composition (RQ2)

To assess the composition of the samples and determine how well they resemble the population, multiple indicators were used, including sex, age, education, employment status, citizenship, household size, and country of birth. These indicators were compared with both the ESS Round 9 (R9) data from Austria (based on face-to-face data collection) and Austrian population statistics.

5.3 Data Quality (RQ3)

Multiple indicators were used to evaluate the data across quality conditions: item-nonresponse, non-differentiation, validity, and internal consistency. We also cover willingness to be recontacted for further research in this section. While this may not strictly be a measure of data quality, we consider it an important metric to compare between conditions. For example, lower agreement to recontact among those completing the longer questionnaire may suggest a more negative respondent experience, which may have implications for future survey participation more generally.

1. Item non-response: To explore potential differences in item non-response between length conditions, a count variable denoting the number of questions respondents skipped throughout the questionnaire was created. For this, we used 121 close-ended, single-choice questions asked to all respondents in both length versions.
2. Non-differentiation: The 21 items from the Human Values Scale were used to assess non-differentiation. These questions were selected as they measure various orientations, and respondents are not expected to select the same answers (which would be a valid form of non-differentiation, Reuning & Plutzer, 2020). An indicator of non-differentiation was generated by computing the coefficient of variation (CV). This CV is calculated for each respondent as the standard deviation of responses divided by their mean. Higher CV values indicate greater differentiation among responses, while lower values suggest less differentiation.
3. Validity: As an indicator of concurrent validity, we analysed the correlation between two items where a negative association was expected: "Gay men and lesbians should be free to live their own life as they wish" and "If a close family member was a gay man or a lesbian, I would feel ashamed". Both statements were evaluated using 5-point agree-disagree scales.
4. Internal consistency: We assessed the internal consistency of the institutional trust scale, which comprises eight items measuring trust in various institutions: the national parliament, the European Parliament, the United Nations, politicians, political parties, the legal system, the police, and scientists. Each item was rated on an 11-point scale, ranging from "no trust at all" to "complete trust".
5. Willingness to be recontacted: In surveys like the ESS, where a panel is built from the main survey (Bottoni & Fitzgerald, 2021), the ability to recontact respondents is crucial for achieving a sufficient sample size at the recontact stage. We compared the proportion of respondents who agreed to be recontacted for further research between the shorter and longer questionnaire group, with this question being placed at the very end of the survey.

6 Analytic Strategy

The analysis was carried out in three main steps. First, we used unweighted data to examine response rate differences between the length conditions. We observed breakoff rates among those completing the web survey between conditions. We also examined the share of web and paper responses across conditions using chi-squared tests.

Second, we examined whether the demographic composition of the samples differed by condition. To do this, we used ANOVAs for age and years of education and chi-square tests for sex, age groups, citizenship, paid work, and country of birth. We also compared the characteristics of the achieved samples with benchmark data (ESS Round 9, and Labour Market Statistics, Statistics Austria, 2019) using dissimilarity indices (d).

For each demographic, d was computed as the sum of the absolute difference between the proportion of the demographic in the sample and its proportion in the benchmark, divided by half. This index can be interpreted as the proportion of observations that would need to change categories in the experimental conditions to achieve perfect agreement with the benchmark data (Biemer et al., 2018). A d value of ≤ 0.10 typically indicates "good" agreement, while a value of $d \leq 0.05$ denotes "very good" agreement. The primary aim of this analysis is to identify which of the two groups best represents the target population before post-survey adjustments are made. Note that a large d value for a group does not necessarily indicate that the group will produce biased estimates. Even if there is a discrepancy between the sample and the benchmark data, weighting adjustments

should rectify them to an extent. However, the dissimilarity analysis provides insight into the extent to which these post-survey adjustments are needed to reduce potential bias. In addition, post-survey adjustments are not always successful in reducing bias and may increase the standard errors of estimates due to increased weight variation. Therefore, conditions with smaller dissimilarity indexes are preferable (Biemer et al., 2018).

And third, we compared potential differences in data quality between length conditions. Item nonresponse and non-differentiation were compared using ANOVAs, while validity was assessed via correlations. Considering the ordinal nature of the variables (5-point scales), the Spearman correlation coefficient was used. Cronbach's alpha values were computed for the shorter and longer questionnaire groups, and the results were compared using a chi-squared-test. Potential differences in willingness to be recontacted were explored using chi-squared tests. All the analyses were conducted using Stata 16, except for the comparison of alpha values, which was run using the cocron package in R (Diedenhofen, 2016).

7 Results

7.1 Response Rates

Fig. 1 shows the response rate and corresponding 95% confidence interval for each condition. The short condition increased the response rate to the questionnaire by 4.2 points, from 34% when the questionnaire was 50 min to 38% when it was about 35 min. The effect size was small (Cramer's V = –0.04), yet the difference was statistically significant (X^2 = 7.76; p = 0.005).

Among those responding online, there was no significant difference in the breakoff rate between the length conditions (X^2 = 0.21; p = 0.645). The breakoff rate for the longer questionnaire was 12% compared with 13% for the shorter questionnaire.

Approximately three out of four respondents completed the questionnaire online, with no significant differences between the length conditions (X^2 = 0.48, p = 0.490) (see Fig. 2). This suggests that the lengthier questionnaire did not deter respondents from completing it on paper, despite the visible thickness of the instrument (36 pages).

7.2 Sample Composition

Sample composition, summarised in Table 1, was comparable for both length conditions. No significant differences were observed between the samples from the shorter and longer conditions. However, in comparison with the benchmark data, both underrepresented young adults (ages 18–29), while overrepresenting both Austrian citizens and individuals born in Austria.

To further assess the representativity of the samples from the two conditions, we compared their distributions against benchmark data. This benchmark data was sourced from population statistics and weighted ESS R9 distributions. Following the approach from Biemer et al. (2018), we used the dissimilarity index (d) to measure unrepresentativeness.

Fig. 3 shows the values of the dissimilarity index, expressed as percentages for each of the variables. Of these, the largest discrepancies are found for citizenship and country of birth, where approximately 10% of cases would need to change categories in the samples to match the population (for citizenship, d indices are d = 11.0 for the lengthier condition and d = 10.3 for the shorter condition; for country of

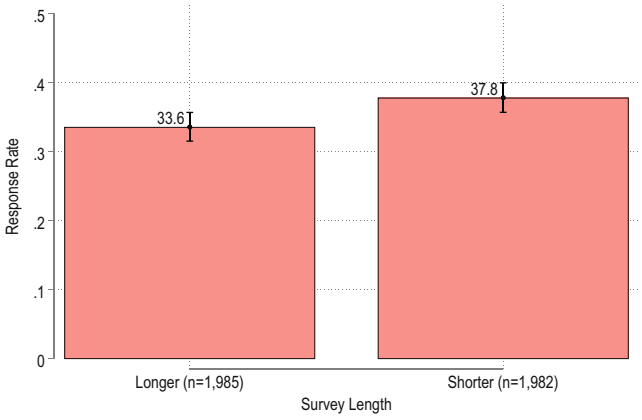

Fig. 1

Response rate by length condition

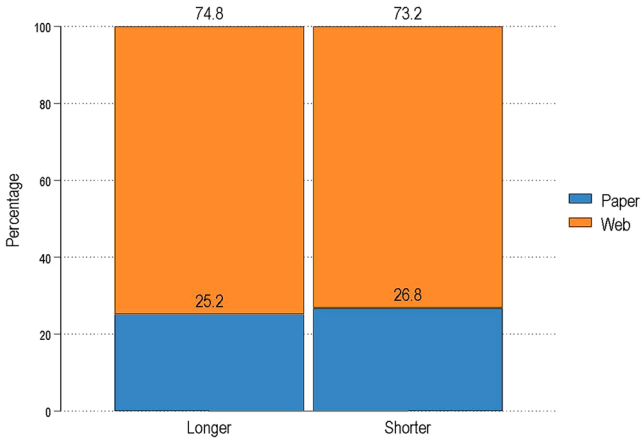

Fig. 2

Share of web and paper response by length condition

Table 1

Sample composition by length condition

Variable	Shorter (35 mins), %	Longer (35 mins), %	t-test/chi sq	Weighted ESS R9 data (2018)	Statistics AT (2019)
Sex					
Female	54	53	0.08	52	51
Male	46	47		48	49
Age					
18–29	11	10	0.59	17	18
30–49	31	32		34	33
50–64	31	31		26	27
65+	28	27		23	23
Average age	52	52	0.07	50	–
Education (Average years)	14	14	–0.35	13	–
In paid work	57	60	0.79	57	50
Citizenship					
Austrian	94	94	0.34	92	83
Non-Austrian	6	6		8	17
Country of birth					
Austria	91	90	0.19	86	80
Other country	9	10		14	20

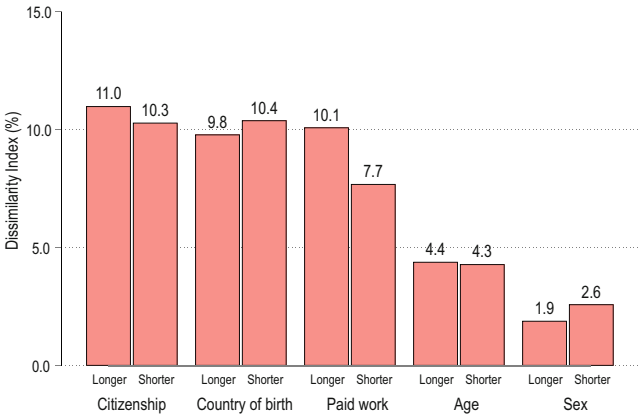

Fig. 3

Dissimilarity indexes for selected items using Austrian population statistics as the gold standard 10–5% dissimilarity is considered to be "good", and 5% or less is considered "very good"

birth, d indices are $d = 9.8$ for the lengthier condition and $d = 10.4$ for the shorter condition).

While both length conditions overrepresented individuals in paid work, the shorter condition fared better, showing slightly closer agreement with the benchmark data. For age and sex, the dissimilarity indices fall in the "very good" range for both conditions.

7.3 Data Quality

The composite variable measuring item nonresponse ranged from 0 to 107, with an average of 3.3 missing observations per respondent ($SD = 9.2$). There were no significant differences in the average number of missing values between the shorter ($M = 3.20$) and the longer conditions ($M = 3.38$) ($t = –0.368$, $p = 0.713$), providing no support to our hypothesis that respondents in the lengthier condition would choose not to answer more questions.

In both short and long versions of the survey, the average number of skipped questions was similar. However, significant differences emerged when comparing the survey modes. The paper-based questionnaires, regardless of their length, showed a significantly higher rate of skipped questions—over twice as many—compared to the web-based questionnaires ($t = 4.977$, $p < 0.001$) as illustrated in Fig. 4. This suggests that while the survey length did not significantly impact response completeness within each mode, the choice of mode (paper vs. web) did.

The Schwartz Human Values Scale (HVS) was used to assess non-differentiation. The HVS scale includes 21 items asking the respondent to assess how alike they are to an

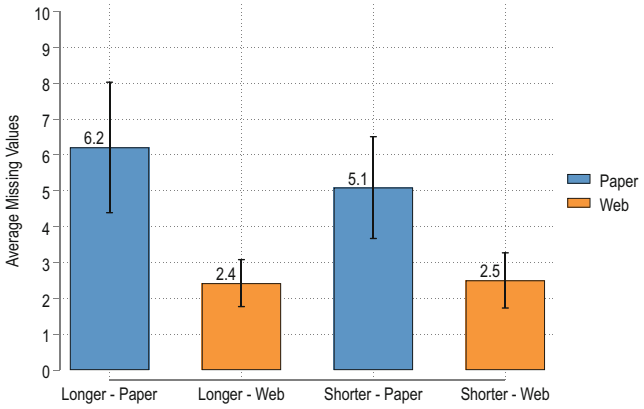

Fig. 4

Average number of skipped questions by length condition and survey mode

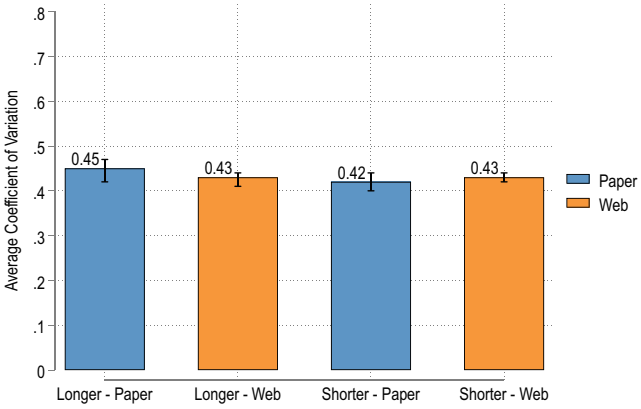

Fig. 5

Average Coefficient of variation by length condition and survey mode

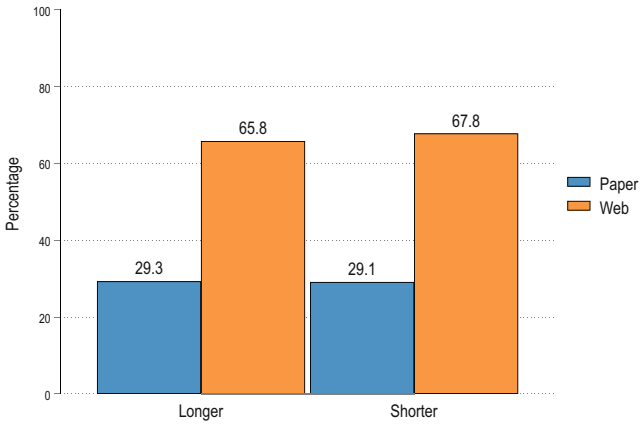

Fig. 6

Willingness to be recontacted by length condition and survey mode

imagined person in their values (e.g., "Being very successful is important to him/her. (S)he hopes that people will recognise his/her achievements"). A 6-point scale, from "very much like me" to "not like me at all" is used, with each question presented individually using a vertical scale in both the web and paper instruments. These items were included in the final section of the questionnaire and asked for both length conditions.

As an indicator of variability, the coefficient of variation was computed (range = 0–1.03; M = 0.43; SD = 0.14), yielding no significant differences between the shorter and longer groups (M = 0.42 vs. M = 0.43, t = –0.960, p = 0.337), or between web and paper (M = 0.43 vs. M = 0.43, t = 0.329, p = 0.742) (see Fig. 5).

As an indicator of concurrent validity, we examined two items for which a negative correlation was expected. Across groups, a negative and moderately strong correlation was found (*rho* = –0.611, p <0.001), with the size of the effect being similar for the shorter and longer conditions (*rho* = –0.584 [–0.641, –0.527] and *rho* = –0.641 [–0.698, –0.584], respectively).

The internal consistency of the institutional trust scale was comparable across both the length versions (X^2 = 0.135, p = 0.713) and the modes (X^2 = 1.529, p = 0.216). The reliability was good to excellent for both the shorter (α = 0.897 for paper and α = 0.893 for web) and longer versions (α = 0.857 for paper and α = 0.900 for web).

Recontact rates were also comparable between the length groups, with slightly more than half of the respondents agreeing to be recontacted for future research (58% in the shorter group and 57% in the longer group, X^2 = 0.138, p = 0.719). However, there were substantial differences between the modes. The proportion of recontact agreements more than doubled in the web compared to paper (67% for web vs. 29% for paper, X^2 = 146.69, p <0.001) (see Fig. 6). This finding is somewhat expected, given that the follow-up was announced to be conducted online, and paper respondents who opted not to respond via the web were likely to have lower propensity to participate online.

8 Discussion

Given the growing prevalence of self-completion methodologies in survey research, it is crucial to understand how design decisions impact survey outcomes in this context. This is particularly important as more surveys that were administered by interviewers transition to self-completion approaches. One key consideration in this transition is the impact of survey length. Lengthy face-to-face surveys may be perceived as being too long to be administered in a self-

completion environment. This may result in researchers feeling forced to cut survey content or adopt matrix designs, both of which result in less comprehensive accessible data. Therefore, it is important to assess the impact of survey length on outcomes in a self-completion environment to ensure that any (sub-optimal) design decisions are based on empirical evidence.

The experiment described above compared a shorter (35 min) and longer (50 min) version of the European Social Survey Round 10 questionnaire in Austria. It was hypothesised that the differential length would affect survey outcomes. This was partly found. A significantly higher response rate was achieved with the shorter version compared to the longer one (38% versus 34%). However, there were no significant differences in sample composition or data quality between the length conditions. This, combined with the relatively small effect size for the response rate difference, suggests that the benefit of reducing the questionnaire length was limited.

Reducing the ESS questionnaire to produce the 'short' version required a substantial reduction in survey content—in this case, removing two rotating modules that cover different topics in each ESS round. Adopting this short version would significantly reduce the amount of data available to users and eliminate the opportunity for researchers to add questions as part of these rotating modules. A decision to shorten the questionnaire therefore comes at a significant cost, but the evidence from this experiment suggests that it does not lead to a large improvement in survey outcomes.

The results of our experiment provide evidence that it is possible to field a relatively long questionnaire in a self-completion environment and achieve reasonable outcomes. However, there are some limitations to what can be inferred from our study, and there are opportunities for further research. First, our experiment was conducted in only one country (Austria), and it is uncertain whether the same results would be seen if the experiment were repeated in other countries. The experience of ESS Round 10, where nine countries adopted a self-completion approach based on the near-full ESS questionnaire, provides some reassurance (Hanson, et al., 2022). Across the nine countries, the median response rate was 34%, with generally good levels of sample representativeness. Nevertheless, the ESS includes a diverse set of 30+ European countries, and it is unclear if length differences may have a greater impact in some countries compared with others. Therefore, carrying out length experiments in other countries would add valuable evidence on this topic. Additionally, running future studies outside the immediate context of the COVID-19 pandemic, which may have impacted on peoples' likelihood to respond, would be beneficial.

Second, the ESS self-completion questionnaire used in this experiment was developed in a short period in response to the COVID-19 pandemic. As a result, the questions were minimally adapted from the face-to-face survey rather than being more extensively 'optimised' for self-administration across web (including smartphones) and paper. Future studies could examine the impact of such optimisation—for example, adapting scales and potentially shortening question wording—to assess how such changes affect survey outcomes.

Third, despite the limited time for developing the ESS's self-completion approach, substantial work was undertaken to prepare the materials for this experiment. This included an earlier three-country self-completion test using a shorter questionnaire, an extensive development phase for the paper questionnaire, including user testing, and consultation with international experts to develop letters and other survey materials following best practices. Such development activities may be more limited for some other studies. There may therefore be a question over whether the length of the questionnaire might play a larger role in likelihood to respond with less extensively developed survey materials.

Assessing with certainty why our survey achieved different outcomes to some of those carried out previously is beyond the scope of this study. However, it is possible that the combination of design factors here—an unconditional incentive coupled with a relatively high conditional incentive, an academically branded survey, a relatively straightforward questionnaire to administer, and materials designed according to Dillman's principles—may not be fully present for other studies. This may partly explain why the evidence from this experiment contradicts findings from some earlier studies, which did find an increased length significantly reduced response rates and data quality. It may also partly explain why other studies more comparable to the ESS, such as the Gender and Generations Survey and the European Values Study, have also achieved promising outcomes with long self-completion questionnaires. Further experimentation with different types of surveys and populations would nevertheless help inform how widely the findings reported here can be generalised beyond the ESS.

Despite these caveats, our findings provide encouraging evidence for the feasibility of using the full ESS questionnaire in a self-completion context. We see the modest drop in response rates observed in the longer version, combined with no evidence of worse sample composition or data quality, as a reasonable trade-off to allow the full ESS questionnaire content to be retained. The outcome of the experiment, combined with further promising evidence on longer questionnaires from the ESS and other studies, has influenced the ESS's decision to retain its full questionnaire once it transitions to a fully self-completion approach in the coming years.

We acknowledge that no two surveys are the same. However, we expect that studies that are carefully designed and exhibit similar features compared to the ESS could also achieve positive outcomes with longer questionnaires, assuming they are as straightforward to administer as the ESS and have similarly low levels of routing. We hope that the results from our experiment may prompt other researchers to at least consider if reducing questionnaire length is always a necessary requirement when shifting from interviewer-administered to self-completion approaches, especially where such reductions will reduce the overall value of the study to its users.

Acknowledgements The data collection for this experiment was overseen and carried out by The Institute for Advanced Studies (IHS) and the Institute for Empirical Social Research (IFES) in Austria. We thank them for their work on this and particularly acknowledge Peter Grand as the ESS National Coordinator in Austria.

We are also forever grateful to the late Don Dillman, whose valuable advice had a major impact both on the design of the experiment and on ESS's self-completion approach more generally. He was a wonderful colleague and critical friend during the entire process of working with the ESS team. His input will impact survey research and the ESS for many years to come. We dedicate this paper to Don's memory.

References

Anduiza, E., & Galais, C. (2017). Answering without reading: IMcs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519.

Beullens, K., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response rates in the European social survey: increasing, decreasing, or a matter of fieldwork efforts? In *Survey methods: insights from the field*.

Biemer, P.P., Murphy, J., Zimmer, S., Berry, C., Deng, G., & Lewis, K. (2018). Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, *6*(2), 240–261.

Bottoni, G., & Fitzgerald, R. (2021). Establishing a baseline: bringing innovation to the evaluation of cross-national probability-based online panels. *Survey Research Methods*, *15*(2), 115–133.

Bradburn, N. (1977). Respondent burden. https://hsrmconference.com/sites/default/files/proceedings/HSRMProceedings02.pdf. Accessed 10.2023. Health Survey Research Methods Proceedings. 49053.

Daikeler, J., Bosnjak, M., & 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.

Daikeler, J., Bach, R.L., Silber, H., & Eckman, S. (2022). Motivated misreporting in smartphone surveys. *Social Science Computer Review*, *40*(1), 95–107.

Diedenhofen, B. (2016). *cocron: statistical comparisons of two or more Alpha coefficients (version 1.0-1)*

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method* (4th edn.). Hoboken: Wiley.

Emery, T., Cabaco, S., Fadel, L., Lugtig, P., Toepoel, V., Schumann, A., Lück, D., & Bujard, M. (2023). Breakoffs in an hour-long, online survey. *Survey Practice*. https://doi.org/10.31235/osf.io/ja8k4.

Enders, C.K. (2010). *Applied missing data analysis*. New York: Guilford.

ESS ERIC (2022). European social survey round 10 questionnaire. https://www.europeansocialsurvey.org/methodology/ess_methodology/source_questionnaire/

ESS ERIC (2021). *ESS10—Austria self-completion experiment (data set). Sikt—Norwegian agency for shared services in education and research. Data set available for researchers to request under special licence*

Fitzgerald, R. (2021). Responding to the pandemic: a 3-country self-completion push-to-web experiment. https://www.youtube.com/watch?v=1zyS3WU3tjE ESS-City-NatCen Survey Methodology Seminar Series.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Groves, R.M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189.

Groves, R.M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *Public Opinion Quarterly*, *64*(3), 299–308.

Gummer, T., & Roßmann, J. (2015). Explaining interview duration in web surveys: a multilevel approach. *Social Science Computer Review*, *33*(2), 217–234.

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

Hanson, T., Fitzgerald, R., Ghirelli, N., & O'Muircheartaigh, S. (2022). *Delivering the European Social Survey during COVID-19: reflections and future implications*. Chicago: AAPOR Annual Conference.

Kotnarowski, M. (2023). *Implementing self-completion mode—experiences of the Polish ESS*

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the European social survey. *Journal of Official Statistics*, *33*(2), 409–426.

Loosveldt, G., & Wuyts, C. (2020). A comparison of different approaches to examining whether interviewer effects tend to vary across different subgroups of respondents. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 311–322). New York: Chapman and Hall/CRC.

Lugtig, P., Toepoel, V., Emery, T., Cabaço, S. L. F., 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*. https://doi.org/10.31235/osf.io/mu8jy.

Luijkx, R., Jonsdottir, G. A., Gummer, T., Stahli, M. E., Frederiksen, M., Ketola, K., Reeskens, T., Brislinger, E., Christmann, P., Gunnarsson, S. T., Hjaltason, A. B., Joye, D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Pollien, A., Sapin, M., Solanes, I., Verhoeven, S., & Wolf, C. (2021). The European values study 2017: on the way to the future using mixed-modes. *European Sociological Review*, *37*(2), 330–347.

Luiten, A., Hox, J. J. C. M., & de Leeuw, E. D. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469–487.

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: a field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*(3), 372–383.

Mavletova, A., & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In D. Toninelli, R. Pinter & P. de Pedraza (Eds.), *Mobile research methods: opportunities and challenges of mobile research methodologies* (pp. 81–98). London: Ubiquity Press.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., McCarthy, J. S., O'Brien, E., Opsomer, J. D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., & Wagner, J. (2020). Transitions from telephone surveys to self-administered and mixed-mode surveys: aAPOR task force report. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smz062.

O'Muircheartaigh, S., Hanson, T., & Fitzgerald, R. (2023). *Challenges and successes of changing mode in a cross-national context: developing a self-completion approach for the European social survey*

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97.

Peytchev, A., & Peytcheva, E. (2017). Reduction of measurement error due to survey length: evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361–368.

Peytchev, A., Peytcheva, E., Conzelmann, J. G., Wilson, A., & Wine, J. (2020). Modular survey design: experimental manipulation of survey length and monetary incentive structure. *Journal of Survey Statistics and Methodology*, *8*(2), 370–384.

Reuning, K., & Plutzer, E. (2020). Valid vs. invalid straightlining: the complex relationship between straightlining and data quality. *Survey Research Methods*, *14*(5), 439–459.

Revilla, M., & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538–545.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108.

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., & Ogilvie, D. (2011). Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: randomised controlled trial. *BMC Medical Research Methodology*. https://doi.org/10.1186/1471-2288-11-62.

Singer, E. (2011). Toward a benefit-cost theory of survey participation: evidence, further tests, and implications. *Journal of Official Statistics*, *27*(2), 379–392.

Smyth, J. D., Olson, K., & Stange, M. (2019). Within-household selection methods: a critical review and experimental examination. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. de Leeuw & B. West (Eds.), *Experimental methods in survey research* (pp. 23–45). Hoboken: Wiley.

Statistics Austria. (2019). Register-based Labour Market Statistics 2019 – Employment Data (Reference date: 31 October). Compiled on 28 June 2021. https://www.statistik.at/fileadmin/pages/54/LabourMarketStatisticsEmployment2019.ods

Verhoeven, S., & Wolf, C. (2021). The European values study 2017: on the way to the future using mixed-modes. *European Sociological Review*, *37*(2), 330–347.

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

Yan, T., & Williams, D. (2022). Response burden—review and conceptual framework. *Journal of Official Statistics*, *38*(4), 939–961.

Yan, T., Conrad, F.G., Tourangeau, R., & Couper, M.P. (2011). Should I stay or should I go: the effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research*, *23*(2), 131–147.