SSOAR

Open Access Repository
www.ssoar.info

Using a Responsive Survey Design to Innovate Self-Administered Mixed-Mode Surveys

Gummer, Tobias; Christmann, Pablo; Verhoeven, Sascha; Wolf, Christof

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 491156185 / Funded by the German Research Foundation (DFG) - Project number 491156185

Empfohlene Zitierung / Suggested Citation:
Gummer, T., Christmann, P., Verhoeven, S., & Wolf, C. (2022). Using a Responsive Survey Design to Innovate Self-Administered Mixed-Mode Surveys. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *185*(3), 916-932. https://doi.org/10.1111/rssa.12835

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY-NC-ND Lizenz (Namensnennung-Nicht-kommerziell-Keine Bearbeitung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

Terms of use:
This document is made available under a CC BY-NC-ND Licence (Attribution-Non Comercial-NoDerivatives). For more Information see:
https://creativecommons.org/licenses/by-nc-nd/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-86014-7

Received: 30 September 2020 | Accepted: 12 February 2022

DOI: 10.1111/rssa.12835

ORIGINAL ARTICLE

Journal of the Royal Statistical Society Series A Statistics in Society

Using a responsive survey design to innovate self-administered mixed-mode surveys

Tobias Gummer[1,2] | **Pablo Christmann**[1] | **Sascha Verhoeven**[1,3] | **Christof Wolf**[1,2]

[1]GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany

[2]School of Social Sciences, University of Mannheim, Mannheim, Germany

[3]Office of City Development and Statistics, City of Heidelberg, Heidelberg, Germany

Correspondence
Tobias Gummer, GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany.
Email: tobias.gummer@gesis.org

Abstract

Implementing innovations in surveys often results in uncertainty concerning how different design decisions will affect key performance indicators such as response rates, nonresponse bias, or survey costs. Thus, responsive survey designs have been developed to better cope with such situations. In the present study, we propose a responsive survey design that relies on experimentation in the earlier phases of the survey to decide between different design choices of which—prior to data collection—their impact on performance indicators is uncertain. We applied this design to the European Values Study 2017/2018 in Germany that advanced its general social survey-type design away from the traditional face-to-face mode to self-administered modes. These design changes resulted in uncertainty as to how different incentive strategies and mode choice sequences would affect response rates, nonresponse bias, and survey costs. We illustrate the application and operation of the proposed responsive survey design, as well as an efficiency issue that accompanies it. We also compare the performance of the responsive survey design to a traditional survey design that would

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

have kept all design characteristics static during the field period.

KEYWORDS

mixed-mode, responsive survey design, survey costs, survey error, survey experiments

1 | INTRODUCTION

Responsive survey designs (RSDs) have been developed to enable more flexible and dynamic fieldwork protocols (Groves & Heeringa, 2006; Tourangeau et al., 2017). For the purpose of our study, we define such designs as those that rely on multiple phases with the gross sample being split and allocated across these phases. In earlier phases, different design characteristics are evaluated based on performance indicators of interest such as response rates, nonresponse bias, or survey costs. In later phases, the survey design is adjusted based on the previous lessons learned. The aim of these adjustments is to improve the indicators of interest (e.g. increase in response rates and/or decrease in survey costs). For instance, based on a first phase of a survey that involves only parts of the gross sample, one might learn that prepaid incentives outperform postpaid incentives with respect to response rates. Considering this insight, and with an aim to increase response rates, the next phase of the survey would change the survey protocols so that all the remaining respondents receive only prepaid incentives instead of postpaid incentives.

In line with other research (e.g. Groves & Heeringa, 2006; Schouten et al., 2018; Tourangeau et al., 2017), we argue that RSDs can be used to implement innovative methods in a survey for which uncertainty exists about how these methods will perform with respect to performance indicators of interest. Advancing general social surveys to web-based data collection can be a challenging situation. Traditionally, these large-scale population surveys were based on interviewer-administered surveying, but more recently researchers doing these kinds of surveys have been under pressure to adopt innovative methods due to increasing nonresponse rates and survey costs (e.g. Luijkx et al., 2021; Wolf et al., 2021). As Luijkx et al. (2021) have argued, knowledge about how a particular design decision will impact a survey's performance indicators is lacking in many countries, which implies that testing to gather more evidence is necessary. If researchers want, or are forced, to implement innovative methods such as web-based interviewing for general social surveys, before this necessary evidence is gathered, they will be operating with uncertainty regarding how each design decision may affect performance indicators—a situation for which RSDs offer a solution.

Unfortunately, despite the growing literature on RSDs, our knowledge is limited with respect to the practical implementation of RSDs to conduct surveys, especially if they must include methods that are innovative, even though the empirical evidence is lacking as to how these methods might impact performance indicators such as response rates or nonresponse bias. In our view, this gap is the result of the dynamic and flexible nature of RSDs, which results in rather unique survey designs that are all considered to be *responsive* survey designs. In other words, RSDs are a class of designs rather than an explicit blueprint on how to design a survey. Thus, a need exists to study specific implementations of RSDs, to carry out more research on how to implement RSDs in practice, and to understand how they perform in comparison to traditional survey designs with static characteristics.

Regarding these research opportunities, we aim to accomplish the following goals with our study. First, we propose an RSD that draws on experimentation for implementing innovative methods. The proposed design relies on experiments in the earlier phases of a survey that can provide information on how to adjust design characteristics in the later phases of the survey with respect to performance indicators of interest (in our case, response rates, nonresponse bias, survey costs). Second, we want to evaluate how the RSD performs in comparison to a traditional survey design with static design characteristics. This comparison, in our view, is required to showcase the benefits of RSDs for social science research that aims to implement innovative methods, especially since RSDs entail 'learning phases' as a means to mitigate risks when the effects of design decisions on performance indicators are unknown.

To answer our research questions, we draw on the European Values Study 2017/2018 in Germany that featured a self-administered mixed-mode survey (mail and web) in which we implemented the proposed RSD. When designing this survey, we aimed at innovating the traditional general social survey-type design (i.e. face-to-face mode) by moving to self-administered modes. However, in our case, we were lacking empirical evidence about how different mode choice sequences and incentive strategies would perform for such a survey in Germany with respect to response rates, nonresponse bias, and survey costs. Therefore, we implemented an RSD with two phases, each having 50% of the gross sample randomly assigned to it. In the first phase, we ran experiments with respect to a mode choice sequence (simultaneous vs. sequential) and incentive strategy (prepaid vs. postpaid). Then, we halted our fieldwork for one week during which time we evaluated each design decision with respect to response rates, survey costs, and the risk of nonresponse based on information from population registers. We used the best performing design to conduct the remaining interviews in the second phase, and we aggregated all the data collected across the two phases in a final data set.

Our study is structured as follows. In the next section, we briefly introduce RSDs and propose a specific RSD that is aimed at implementing innovative design characteristics. Then, we detail our case study—the European Values Study 2017/2018 in Germany—especially its RSD, which was used to implement self-administered modes (mail and web-based surveying) in a general social survey. Next, we present our findings, and then close with concluding remarks, practical suggestions concerning the application of RSDs, and an outlook for future research opportunities.

2 | RESPONSIVE SURVEY DESIGN FOR IMPLEMENTING INNOVATIONS

RSDs rely on multiple phases of data collection. Groves and Heeringa (2006) have summarized four steps that RSDs incorporate: '(a) preidentify a set of design features potentially affecting costs and errors of survey estimates, (b) identify a set of indicators of the cost and error properties of those features and monitor those indicators in initial phases of data collection, (c) alter the features of the survey in subsequent phases based on cost-error trade-off decision rules and (d) combine data from the separate design phases into a single estimator' (p. 440). In other words, the use of different design characteristics with respect to indicators of interest is tested in earlier phases and some of them are then implemented in later phases of a survey. An important feature distinguishing an RSD from a survey that uses pre-testing is that an RSD collects data across all phases and combines them in a final data set (step d).

The adaptive design paradigm (Wagner, 2008) emerged in close proximity to RSDs. The labels adaptive survey design (ASD) and RSD often are used synonymously and are not mutually exclusive (Schouten et al., 2018, pp. 19–26). In the present study, we refer to RSDs as surveys that collect data across multiple phases, adjust survey protocols based on prior experiences, and integrate data collected across all phases in a final data set. To clarify, in the present study, we focus on a design that treats parts of a sample differently, depending on the phase of the survey to which the parts have been allocated in the beginning. We do not focus on what we understand to be an ASD in which the strata of a sample are treated differently, for example, groups at risk of nonresponding. We refer readers who are interested in definitions, distinctions, and combinations of ASDs and RSDs to the excellent overview provided by Schouten et al. (2018).

Groves and Heeringa (2006) have argued that RSDs can be used in situations in which uncertainty exits as to how specific design decisions would affect survey errors and costs (for a similar argument, see Schouten et al., 2018). Accordingly, in applications of RSDs, insights from earlier phases of a survey are used to improve the design in later phases (e.g. Axinn et al., 2011, 2021). For instance, Axinn et al. (2021) used an RSD for a campus student survey to investigate how different incentive amounts ($15 or $30) and different recruitment strategies performed with respect to sample composition and substantive measures. In a similar situation, an RSD was used in the 2012/2014 Beginning Postsecondary Students Longitudinal Study in the U.S. to determine the baseline incentive amount (Hill et al., 2016). In this case, the 'calibration' sample was split into 11 groups, and the incentive amount (provided as a cheque) varied between $0 and $50. The study found that $30 was the optimal amount for the main sample. We concur that RSDs are well suited to implement innovative survey design characteristics in situations in which it is uncertain how they will impact relevant performance indicators such as response rates, nonresponse bias, or costs.

For the purpose of implementing innovative design characteristics in a survey, we propose an RSD that relies on experimentation in earlier phases of a survey to learn about and adapt the best performing design characteristics in later phases. Thus, we define an RSD as consisting of k design phases p_k, where $k > 1$. Accordingly, a survey's gross sample S is split into k subsamples ($s_1 \ \dots \ s_k$), each of sample size n_k. Each subsample is allocated to one phase of the survey (i.e. s_1 allocated to p_1). However, subsamples do not have to be equally sized, and thus the phase allocation probabilities ($\pi_1 \ \dots \ \pi_k$) may vary. In phase p_j (where $j < k$), l different design characteristics d_l are experimentally tested with respect to i indicators q_i that are calculated for each design characteristic (q_{ijl}). For this purpose, s_j is randomly split into l subsamples, and each is allocated to an experimental condition that implements the respective design characteristic d_l. If experimental groups are equally sized, then $n_{jl} = \frac{n_j}{l}$. Again, experimental groups may differ in their sample size when allocation probabilities (π_{jl}) vary between them. The reason for varying sizes of subsamples for phases and experimental groups may be that prior knowledge exists for some design characteristics compared to others.

We assume that it is uncertain which values q_{ijl} will take prior to data collection. Although response rate, nonresponse bias, and survey costs have been used in prior studies as performance indicators (e.g. Axinn et al., 2011; Groves & Heeringa, 2006), we argue that other sources of error (for an overview, see the Total Survey Error Framework; Groves et al., 2009; Weisberg, 2005) or survey operations (e.g. the share of participants in a specific mode, number of required contacts until participation, field duration) might be of interest to researchers. After data collection for p_j is complete, q_{ijl} are calculated, and it is determined which design characteristic performed best with respect to q_{ijl}. If $i > 1$, this decision can take the form of a cost-error trade-off (see above, step c). The selected design characteristics are then implemented in p_{j+1}. After data collection is complete

for the last phase of the survey, data from all samples are combined and constitute the final data set that is used for substantive analyses. In the least complex variant, this RSD is operated with $k = 2$, where in p_1 all experiments are conducted and p_2 is run with the best performing design characteristics.

In contrast to RSDs (and ASDs) are static survey designs. Static survey designs resemble the traditional way of fielding surveys—deciding on all the design characteristics prior to the fieldwork start and then sticking with them until the fieldwork is finished. In these cases, survey design characteristics are held constant and are not adjusted, and the whole sample is treated in the same way (note that this characteristic also distinguishes static survey designs from what often is referred to as 'static ASDs').

Our proposed RSD collects data for a sample across k phases. Across phases, knowledge is gathered on how to improve survey protocols via experimentation. Based on this knowledge, protocols in subsequent phases are adjusted to improve the relevant outcome indicators (e.g. response rates, nonresponse bias, survey costs). The final data set contains all data collected across all phases. Consequently, the proposed RSD has an efficiency issue: the final data set includes samples that consist of data that are collected under non-optimal design characteristics and score lower on q_{ijl}. In other words, parts of the sample are allocated to phases that are used to 'learn' and thus (partially) employ underperforming survey protocols. To improve the overall performance, n_k in the first phase(s) should be as small as possible and as large as necessary to yield valid results to inform the next phase(s) of fieldwork. Although the goal is to minimize the number of respondents allocated to these learning phases, at the same time, enough respondents (and resources) need to be allocated to the 'learning phases' to enable insights.

3 | THE RESPONSIVE SURVEY DESIGN OF THE GERMAN PART OF THE EUROPEAN VALUES STUDY 2017/2018

We implemented the proposed RSD in a self-administered mixed-mode survey that was fielded as part of the EVS 2017/2018 in Germany. Surprisingly, at the time of designing the survey, little was known about probability-based self-administered mixed-mode (mail and web) surveying for general social surveys in Germany and how basic design decisions might impact response rates. When considering the first steps of contacting respondents, we already were facing a lack of empirical evidence on how different implementations would perform with respect to two important characteristics: first, the sequence in which mode choices would be offered to a respondent when sending them a survey invitation, and, second, how best to incentivize respondents for their participation.

In Germany, self-administered modes of probability-based samples are used mostly in the context of panel surveys that rely on face-to-face recruitment interviews, and then switch to self-administered modes for re-interviews (Blom et al., 2015; Bosnjak et al., 2018). Prior to the start of our data collection, only one contemporary study by Mauz et al. (2018) was known to us, in which the authors fielded a mixed-mode survey of the general population in Germany based on a random sample with self-administered elements (web, mail, telephone). Unfortunately, these authors did not attempt to gather knowledge on how to design basic fieldwork protocols for a self-administered mixed-mode setting. For example, Mauz et al. (2018) did not offer incentives or test different incentive strategies, despite strong evidence that monetary rewards substantially increase survey participation (e.g. Church, 1993; Felderer et al., 2018; Pforr et al., 2015; Singer & Ye, 2013).

Although additional research was available on how to implement self-administered mixed-mode surveys, at the time of designing our survey, we were skeptical whether the results of these studies could simply be transferred to the German setting. Moreover, the findings from several studies have indicated that the survey climate is quite distinct in different countries. The European Social Survey (ESS), well-known for its rigid standards for cross-national implementation, has yielded significant differences in response rates between countries (Stoop et al., 2010). The ESS even conducted tests on the implementation of mixed-modes surveys of which Villar and Fitzgerald (2017) reported variations in how these surveys performed in different countries. With respect to the eight countries under investigation, face-to-face response rates varied between 46% (Switzerland) to 70% (Poland). Four countries also tested telephone interviews and reported response rates ranging from 18% (Hungary) to 38% (Switzerland). The other four countries employed various mixed-mode designs that combined face-to-face interviewing with either web or telephone interviews and had response rates ranging from 38% (Sweden: face-to-face and telephone) to 64% (Estonia: web and face-to-face).

de Leeuw et al. (2018) also found between-country differences in response rates by using the data of the Labour Force Survey that covered 27 countries across a period of 36 years (1980–2015). Regarding response rates, the authors estimated an intra class correlation (ICC) coefficient of 0.88, which indicated that 88% of the variance could be attributed to differences between countries and 12% to a variation within countries over time. Large differences in response rates also have been reported when comparing the general social surveys of the U.S. and Germany (e.g. Gummer, 2019) where face-to-face response rates for the U.S. slightly decreased from about 75% in the 1980s to about 70% in the 2010s, and response rates in Germany declined from about 70% in the 1980s to about 35% in the 2010s. In addition, a meta study by Daikeler et al. (2020) found countries to moderate the relationship between survey modes and response rates. For instance, these authors reported web survey response rates in the U.S. to be only 9 percentage points lower than other survey modes, whereas in Great Britain and the Netherlands this difference was 16 percentage points.

The EVS 2017/2018 in Germany featured three distinct surveys: a face-to-face survey, a self-administered mixed-mode survey with a static survey design, and a self-administered mixed-mode survey using our proposed RSD. Regarding the EVS Germany, one probability-based sample was drawn from German municipalities' population registers, which was then randomly assigned to the face-to-face and both mixed-mode surveys. In addition, each survey was conducted independently from the others. In the present study, we focused on the self-administered mixed-mode survey that relied on our proposed RSD in its least complex form of $k = 2$. Whereas the face-to-face survey and the static design mixed-mode survey featured the full EVS questionnaire with a mean length of ∼59 min, we applied a split questionnaire design (Peytchev & Peytcheva, 2017; Raghunathan & Grizzle, 1995) to reduce the questionnaire length for the self-administered mixed-mode survey (featuring the RSD) to ∼38 min (mean for the web mode). For the split questionnaire design, the full questionnaire was split into a core module and four modules with further questions. Based on these modules, six different questionnaire versions were developed, each of which contained the same core questions and two out of the four additional modules of questions. We randomly allocated respondents to one of the questionnaire versions. In the present study, we focused solely on the self-administered survey that featured an RSD, since this survey was subject to the aforementioned uncertainty with respect to how response rates, nonresponse bias, and survey costs could vary regarding certain design characteristics. We refer readers who are interested in more general findings regarding the experimentation in the latest EVS to the following studies: Luijkx et al.

(2021) introduce the EVS data set releases, detail the split questionnaire design, and discuss the experiments carried out in six countries from a cross-national perspective; and regarding Germany, Wolf et al. (2021) compare the EVS face-to-face and self-administered mixed-mode surveys.

3.1 | Responsive survey design

Regarding the mixed-mode survey (featuring the RSD), we aimed at collecting around 3,000 interviews of net cases. We randomly divided the mixed-mode gross sample into two similar-sized subsamples (s_1, s_2, with n_1 and $n_2 = 6{,}480$): one to be fielded in the first phase (p_1) and the other to be fielded in the second phase (p_2) of the RSD. However, in practice, we did not use the whole s_2 because the survey was more successful in p_1 than we previously anticipated (see Section 4.1). So only $n_2 = 2{,}916$ randomly selected addresses were fielded in p_2. Furthermore, we randomized the respondents of s_1 into groups in the first phase (see Figure 1). In addition, for the first phase, we randomly and independently allocated respondents to experiments concerning incentive strategies (5€ prepaid vs. 10€ postpaid) and mode choice sequence (sequential vs. simultaneous). Given that the German general population showed lower levels of internet penetration and smartphone adaption compared to other advanced economies (Poushter, 2016; Taylor & Silver, 2019), we implemented the experimental variation of mode choice only for respondents aged 18–59 years (about two thirds of the gross sample). We applied the second experimental condition—incentive strategies (5€ prepaid vs. 10€ postpaid)—equally to all age groups. Thus, our design resulted in six experimental groups (see Figure 1). We randomly allocated respondents to each group (1,100 gross cases each) to be able to draw meaningful conclusions from the comparison of the different groups.

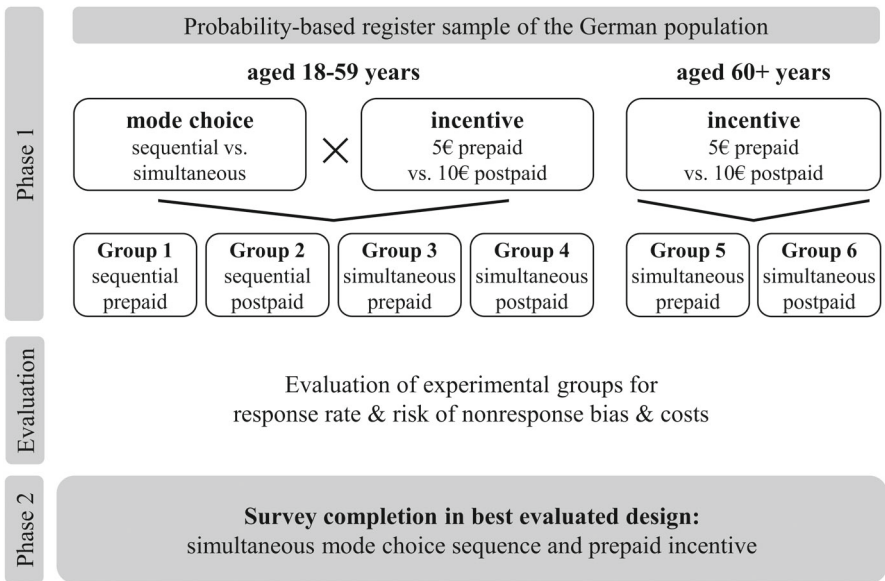

FIGURE 1 Responsive self-administered mixed-mode survey design in the European Values Study 2017/2018 in Germany

With respect to the sequential and simultaneous mode choice conditions, we varied how respondents were given the possibility to participate via web or a mailed questionnaire. Figure 2 illustrates both mode choice sequences with three contact attempts each. For the sequential condition, on the first and second contact attempts, we provided respondents with a link to participate via a web questionnaire (i.e. push-to-web), and on the third contact attempt, we delivered a mailed paper questionnaire to the respondents. In contrast, regarding the simultaneous (i.e. concurrent) condition, we provided respondents with mailed questionnaires as well as web participation links on the first contact attempt. Apart from the timing of providing the different possibilities for participation, both conditions were identical.

The initial contact attempt for phase 1 was made on November 16, 2017 (the date of dispatch). One week later (November 23, 2017), we sent reminder letters (2nd contact) to all persons irrespective of whether they already had participated. Two weeks later (December 7, 2017), we sent a second reminder (3rd contact) to all persons who had not already participated by that date. This reminder also included a paper-based questionnaire for all respondents in the sequential mode choice condition or a second paper-based questionnaire for all respondents in the simultaneous mode choice condition. In the first week of January 2018, we evaluated the outcome and performance of the first phase (for more details see Section 3.2). Until that date, 1,934 respondents had participated in the survey, which exceeded our expectations and left us with a target-*N* of 1,066 for phase 2. Based on an evaluation of the described experiments, we decided to select the best-performing combination of design choices (i.e. the simultaneous mode choice with a 5€ prepaid incentive, see below) for the phase 2 implementation. Since all the necessary routines for phase 2 already had been developed and implemented by the fieldwork institute as part of phase 1, fieldwork for phase 2 started on January 25, 2018, shortly after we completed the evaluation phase. For phase 2, we used survey materials (i.e. invitation letters, respondent information, etc.) similar to those used in phase 1, and we sent two reminders with the same intervals between them.

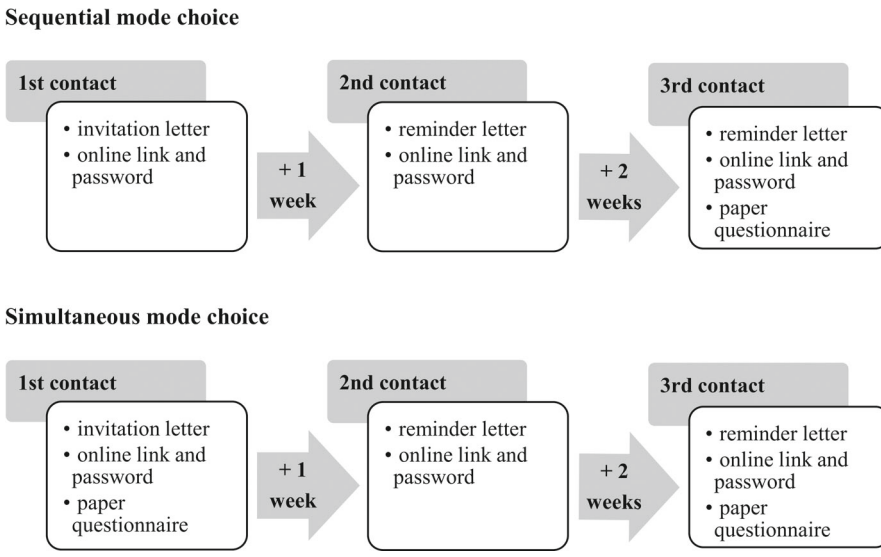

FIGURE 2 Sequential and simultaneous mode choice strategies in the European Values Study 2017/2018 in Germany

For the final data set, we pooled all the data collected in phases 1 and 2 (N = 3,105). Thus, we integrated the data across all experimental groups in phase 1 and the data collected in phase 2 in the same data set. Phase 1 data amounts to 62.3% of this final data set.

After we made the contact and incentive decisions for phase 2, we contacted (for a fourth time) all the respondents of phase 1 who had not yet participated. We did this to gather insights for future surveys on the effects of an additional fourth contact six weeks after the third contact. In total, this fourth contact added 170 additional interviews to phase 1 (8% of all the interviews). However, since we did not intend to implement this design feature in phase 2 (for this phase, we contacted respondents only three times), we omitted from our analyses all the phase 1 interviews collected after the third reminder so to enhance the comparability of the phases.

3.2 | Evaluation of outcome and performance of phase 1

To evaluate the outcome and performance of phase 1, we focused on key indicators that we could analyse within a short amount of time (i.e. the first week of January) and that were most important to us regarding conducting the EVS in Germany. First, as an indicator for survey outcome during ongoing fieldwork, we calculated response rate 6 (RR6) as defined by the American Association for Public Opinion Research (AAPOR) standard definitions (AAPOR, 2016).

Second, to indicate the balance of the sample with respect to the socio-demographic characteristics from population registers and, thus, the risk of nonresponse bias (Schouten et al., 2016), we calculated the adjusted coefficient of variation (CV) of response propensities as suggested by Schouten et al. (2009). For this calculation, we fitted nonresponse models to predict the response propensities for each person in the gross sample based on the information we had from the population register (age, sex, citizenship, urbanity, East-West Germany). The CV is defined as the standard deviation of a sample's response propensities divided by the response rate. A value close to zero indicates a higher sample balance and, thus, a lower risk of nonresponse bias, whereas a higher value indicates a less balanced sample. We obtained the CVs and their approximated standard errors using the R-code provided by the Representative Indicators for Survey Quality (RISQE) project (de Heij et al., 2015).

Third, as an indicator for survey costs, we estimated the costs of collecting the remaining 1,066 interviews, given the different design decisions available to us. These cost estimates relied on the phase 1 outcomes at the time of evaluation. Thus, we drew on the price quotes from the institute tasked with fielding the survey. Additionally, based on the response rates and prices, we calculated the costs of materials (postage, paper, and printing). When comparing costs by survey design, we used survey completion with the experimental condition '5€ prepaid incentive and simultaneous mode choice' as a reference, and calculated cost differences (in percentages) of the other conditions.

4 | RESULTS

In the following subsections, first, we report survey outcomes with respect to the key performance indicators from the learning phase of the RSD (phase 1). Second, we provide the results of adjusting the survey design in phase 2. In the third subsection, we compare the performance of the

RSD to hypothetical scenarios of running static survey designs. Then, we illustrate the previously highlighted efficiency issue of our proposed RSD.

4.1 | Operating the responsive survey design: Learning from phase 1

Figure 3 shows the differences in the response rates of the six experimental groups on the date of evaluation in the first week of January 2018. When comparing all prepaid to postpaid groups, we consistently found that response rates where highest among the respondents who received an unconditional incentive. This finding is consistent with prior research on incentives used in face-to-face surveys in Germany (e.g. Pforr et al., 2015).

With respect to the main effect of mode choice sequence (not shown in Figure 3), our data showed a significant higher response rate for the simultaneous mode choice sequence (RR6 = 0.30) compared to the sequential mode choice sequence (RR6 = 0.27) for the 18–59 years old respondents, when performing a two-sample test of proportions ($p = 0.035$). As Figure 3 suggests, this effect appears to be conditional on the incentive type provided. Although no significant difference seems to exist between the sequential (RR6 = 0.22) and the simultaneous mode choice sequence (RR6 = 0.23) with regard to the 10€ postpaid condition ($p = 0.555$), the difference is significant regarding the 5€ prepaid condition ($p = 0.016$). This finding indicates that the beneficial effect of the prepaid incentive is most pronounced in the simultaneous design (RR6 = 0.37).

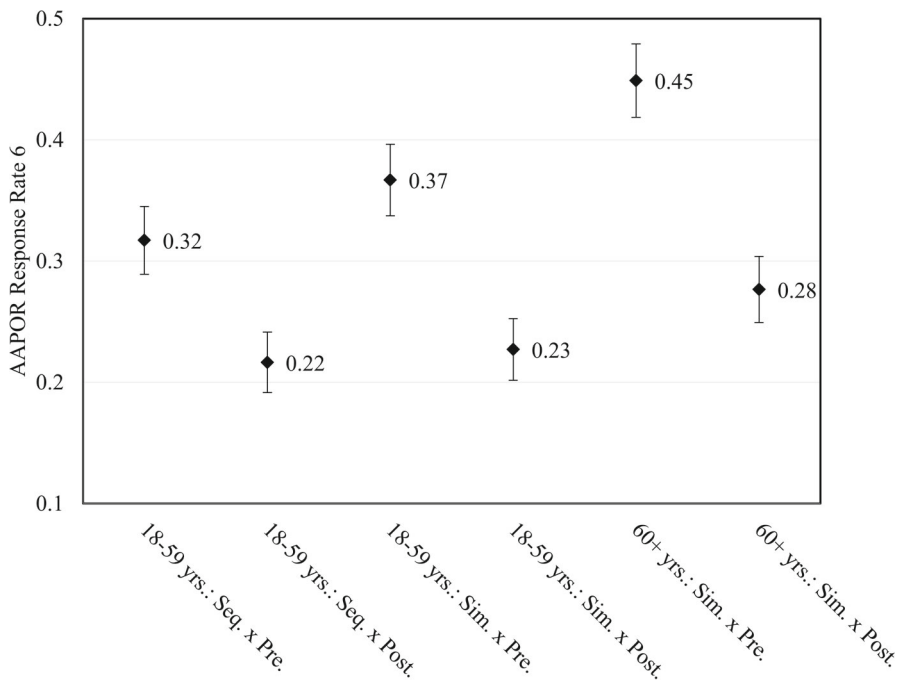

FIGURE 3 Response rate across the experimental groups in phase 1 of the European Values Study 2017/2018 in Germany (at the time point of evaluation)

Note: Seq. = sequential mode choice; Sim. = simultaneous mode choice; Pre. = 5€ prepaid incentive; Post. = 10€ postpaid incentive

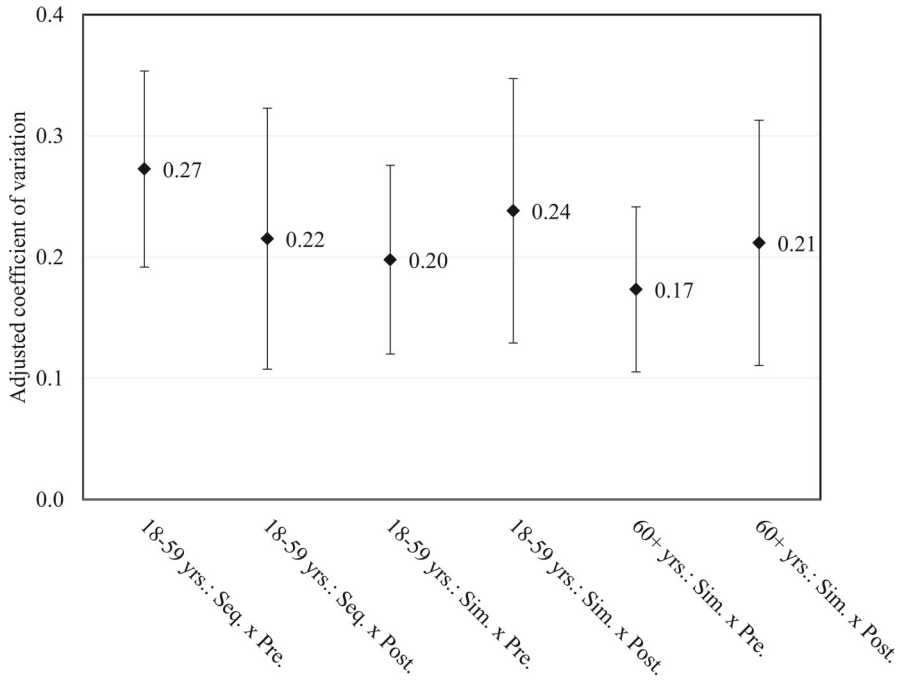

FIGURE 4 Adjusted coefficient of variation across the experimental groups in phase 1 of the European Values Study 2017/2018 in Germany (at the time point of evaluation)

Note: Seq. = sequential mode choice; Sim. = simultaneous mode choice; Pre. = 5€ prepaid incentive; Post. = 10€ postpaid incentive

The adjusted CVs and their 95% confidence intervals for each experimental group are presented in Figure 4. Although we found some differences in the magnitude of the CVs, the confidence intervals overlapped in all instances. These large confidence intervals are likely the result of the low model fit of the nonresponse models that we used as a basis to calculate the CVs and/or the sample size (since variances of variances tend to be larger as variances of means). This phenomenon is common when only register-based information is available for nonresponse prediction. Based on our findings, we conclude that each design decision results in a comparable sample balance and, thus, a similar risk of nonresponse bias based on the variables of the nonresponse model (i.e. age, sex, citizenship, region of residence, urbanity, East/West Germany).

Table 1 shows the two cost indicators (total costs and material costs) that we estimated for the different design choices for phase 2 to complete the survey. Note that we invited all persons aged 60 and older using a simultaneous mode choice sequence. Regarding total survey costs, a combination of sequential mode choice and prepaid incentives seems to be the cheapest option. However, a combination of simultaneous mode choice and prepaid incentives is only slightly more expensive. The most expensive option for survey completion is a combination of simultaneous mode choice and postpaid incentives (Table 1, left column).

Focusing only on material costs (Table 1, right column), we found costs to be higher for the simultaneous mode choices in comparison to the sequential modes because paper questionnaires need to be printed and mailed to all respondents of the gross sample from the beginning. Comparing prepaid and postpaid conditions, a pull-effect of the unconditional incentive can be observed because more respondents already have agreed to participate from the beginning, and so a smaller part of the gross sample needs to be contacted again throughout the survey, which results in lower

TABLE 1 Relative survey costs across the experimental groups in the European Values Study 2017/2018 in Germany

Survey design	Total costs	Costs of material
simultaneous × prepaid	ref	ref
simultaneous × postpaid	113%	167%
sequential × prepaid*	98%	90%
sequential × postpaid*	103%	137%

Note: Ref = reference category. *Respondents aged 60 or older receive a simultaneous (concurrent) mode choice. All figures in percent, relative to the reference category.

material costs. Thus, when focusing only on material costs, the differences between postpaid and prepaid conditions are more pronounced.

Our findings regarding response rates, risk of nonresponse bias, and survey costs enabled us to decide on the design for phase 2 of the survey during January 2018. Based on its positive impact on response rates, we decided in favour of a simultaneous mode choice with a prepaid 5€ incentive. This decision was supported by our finding that this design did not result in a significantly higher risk of nonresponse bias compared to the other designs, and the negligible differences in survey costs. Only a sequential prepaid design would have resulted in slightly lower total costs, however, at the expense of slightly lower response rates.

4.2 | Updated survey protocols in phase 2 of the responsive survey design

Our data indicated that implementing simultaneous mode choice and prepaid incentives in phase 2 of the survey for the remainder of the gross sample performed as expected. Regarding the remaining gross sample for phase 2, we contacted 2,916 persons, which resulted in 1,171 interviews and, thus, achieving our target net sample size of 3,000. The response rate in phase 2 was 42.2%, which was similar to the comparable experimental condition in phase 1 (39.4%, difference not statistically significant with $p > 0.05$). Also, the share of web interviews was comparably low: 17% in phase 1 and 18% in phase 2. Regarding the risk of nonresponse bias, we found overlapping confidence intervals between the respective samples in phase 1 and phase 2 (adjusted CV phase 2 = 0.126, $95\% - CI_{[lower;upper]} = [0.083; 0.169]$) and phase 1 (adjusted CV phase 1 = 0.164, $95\% - CI_{[lower;upper]} = [0.105; 0.223]$). We interpret this finding as indicating comparable risks of nonresponse bias in phase 2 and phase 1.

4.3 | Comparing static survey design to responsive survey design

Based on our previous analyses, Table 2 provides a comparison between the RSD and the hypothetical scenarios for running a static survey design. These hypothetical scenarios resemble situations in which we would have decided to use a sequential or a simultaneous mode choice and prepaid or postpaid incentives before beginning the data collection. Thus, these scenarios do not include an adjustment of survey characteristics across multiple phases. We used data from the six experimental groups in phase 1 to estimate the response rate and the CV for each static survey design. Based on the response rates, we calculated the gross sample sizes necessary to

TABLE 2 Projected outcomes of static survey designs in comparison to the responsive survey design

	Static survey designs				Responsive survey design
Outcome	A* Sequential & prepaid	B* Sequential & postpaid	C Simultaneous & prepaid	D Simultaneous & postpaid	E
Required gross sample	8,330	12,696	7,609	12,322	8,737
RR	36%	24%	39%	24%	34%
Adjusted CV	0.243	0.221	0.164	0.218	0.183
Total costs	98%	103%	100%	113%	ref

Note: Projected estimates based on commercial offers and the study results (phase 1 at evaluation time point and phase 2); design E consists of data collected in phase 1 and phase 2 of the survey; ref = reference category, all figures are based on a net sample of N = 3,000; RR = AAPOR Response Rate 6; CV = adjusted coefficient of variation. *Respondents aged 60 or older receive a simultaneous mode choice sequence.

achieve 3,000 net cases for each static survey design, which enabled us to estimate the costs accordingly. For the RSD, we calculated the final outcome rates by combining all the data from phase 1 and 2. Then, for comparability, we rescaled the RSD's figures to 3,000 net cases.

Our estimations indicated that the theoretically best-performing static survey design would have been C, which showed the best overall performance. Our RSD differed in its performance from the static survey designs. Relative to the hypothetical designs A and C, it underperformed with respect to the outcomes studied, which is the price of uncertainty. Since we did not know how different design choices would perform, we included design features in phase 1 of the RSD that turned out less than optimal. Static survey design C showed an estimated five percentage points higher response rate, and consequently, a smaller gross sample size. Nonetheless, design C would have had the same costs as the RSD, and design A would have produced two percentage points lower total costs. Compared to the hypothetical designs B and D, the RSD clearly performed better with respect to costs, response rate, and necessary gross sample size.

In our view, the previous analyses support the notion that our proposed RSD is a viable way to mitigate the risks of data collection that stem from an uncertainty about which survey outcomes to expect. Of course, had we known at the design phase of the survey what we know now, we could have saved costs and yielded a higher response rate (i.e. an increase of five percentage points). However, at the time, the effects of incentives and the sequencing order for modes were largely unknown for Germany. Thus, these findings illustrate the utility of using RSDs in this particular case and underline the importance of considering the efficiency issue discussed in Section 2.

5 | CONCLUSION

In the present study, we set out to address the lack of research on how to implement and evaluate RSDs in practice. We proposed an RSD that relies on experimentation in earlier survey phases to implement innovative design characteristics for which—pre-data collection—it was uncertain as to how they might impact key survey performance indicators. We illustrated the implementation

of this design based on the German part of the EVS 2017/2018 using response rates, nonresponse bias, and survey costs as our performance indicators of interest. Using an RSD helped us to converge to a design that uses prepaid incentives (instead of postpaid incentives) and a simultaneous mode choice sequence (instead of sequential mode choice sequence). Most importantly, first, we found that prepaid incentives resulted in higher response rates compared to postpaid incentives. Second, we found that combining a simultaneous mode choice sequence with prepaid incentives was only two percentage points more expensive in total costs than using a sequential mode choice sequence with prepaid incentives (which was the cheapest way of designing our particular survey). Third, we found that different combinations of incentives and mode choice sequences did not differ with respect to the risk of nonresponse bias.

Our findings have practical implications for survey research. RSD is a powerful tool to implement new methods in a survey for which sufficient knowledge does not exist as to how these methods will affect key performance indicators such as response rates, nonresponse bias, or survey costs. Typically, innovations are associated with uncertainty, especially if sound research is lacking or if country-specific research is missing and the generalizability of research from other countries is questionable. In our case, our proposed RSD helped us to reduce the risk of selecting underperforming design characteristics with only a small loss in overall performance when advancing a general social survey from a face-to-face mode towards mail- and web-based interviewing. For instance, we found that collecting 3,000 cases by employing an RSD was only two percentage points more expensive than using (by chance) the cheapest static survey design. If relevant knowledge is available, a researcher can select the best possible survey design from the beginning. In most cases, however, there are design characteristics with unknown effects on outcome measures. In these situations, the appeal of an RSD is its ability to converge towards an optimal solution without too much of a decrease in overall performance. Researchers who want to implement our proposed RSD can improve performance indicators by running as few cases as possible in earlier "learning" phases. However, each phase must contain enough respondents to be able to draw conclusions on which adjustments can be made to the design.

With respect to the important question of allocating cases to different phases of an RSD, we would point out that if prior knowledge exists concerning the differences and variations in indicators (e.g. which differences to expect if a prepaid incentive is used instead of a postpaid incentive), power analyses can be used to calculate the sample sizes required to detect an effect of a given size with a specific probability (Cohen, 1988; Murphy et al., 2014). In this case, researchers can pre-define the magnitude of effect sizes that would lead them to consider changing a survey feature. If the necessary information to perform a power analysis is available, researchers can determine the minimum number of cases required in the learning phases of the RSD. This challenge of allocating cases to phases becomes more complex when multiple performance indicators are to be considered, such as those in our study—response rates, nonresponse bias, and survey costs. In these cases, the power analyses should be based on all indicators and the consideration that they might be interrelated.

The limitations and findings of the present study provide opportunities for future research. First, a key characteristic of RSDs is that data are collected across different phases of a survey. Then, the aggregate of these data constitutes the final data set. In the present study, we only briefly touched on the issue of combining data across phases; instead, we just pooled the data in one final dataset. However, given the differences in survey design (incentives, mode choice sequence), it remains an open question as to whether other approaches of fusing data together might help to better mitigate these differences. The results of our study further suggest that the data obtained by each experimental group (i.e. the combination of survey design characteristics) vary with respect

to performance indicators. Depending on how much weight an individual data set gets in the final data set, it may have either a positive or negative impact on the overall data quality of the final data set. The possibility to manage the impact of individual data sets on the final data set is in controlling the allocation probabilities of gross sample cases to phases and experimental groups (i.e. reducing the impact of subsamples that are subject to survey characteristics that are likely to perform less well in the final dataset). We invite future research to advance our knowledge about this key RSD issue.

Second, we operated our RSD based on a pre-selected set of indicators that we wanted to improve. The set of relevant indicators will vary among studies, and other researchers may consider different indicators to be of greater importance (e.g. share of web respondents). In the case of mixed-mode surveys, minimizing mode-specific measurement error and, thus, reducing differences between modes can be important to many other researchers. However, when operating an RSD, it is important to focus on indicators that are available, calculable, and analysable during the field period. We see merit in future research on RSDs that identifies and defines indicators for fieldwork monitoring and data quality, highlights techniques that enable practitioners to improve these indicators, and suggests statistical methods to best evaluate these indicators during the evaluation phase.

Third, our case study is the least complex version of our proposed RSD with only two phases, similar to those discussed in prior studies (e.g. Groves & Heeringa, 2006). However, work that advances RSDs further should extend this limited perspective and cover complex multi-phased RSDs that can provide learning at various stages and via different methods (observation, experimentation). This approach appears to be particularly important for the application of RSDs in long-term panel studies with many waves. More research on this issue is certainly desirable.

Fourth, our implementation of our proposed RSD in the EVS requires replication to test whether our findings with respect to design characteristics are generalizable to other contexts (e.g. the effect of using prepaid incentives on response rates in self-administered mixed-mode surveys). We believe it would be valuable to perform replications in different countries, with different target populations, or different survey topics. We also would highlight that the magnitude of our findings regarding how our RSD performed in comparison to other hypothetical scenarios (see Section 4.3) will likely vary in other studies that rely on a different set of performance indicators, use a different questionnaire, or define different target populations. However, the efficiency issue that is the reason for these differences (see Section 2) will be present in all RSDs that are designed as we propose, and only the magnitude of this issue will vary.

Fifth, the terms RSD and ASD often are used interchangeably, but as Schouten et al. (2018) have argued, responsive and adaptive designs—with their particular strengths and weaknesses—are both specific versions of a more flexible survey design class. To be able to adapt to a fast-changing survey climate and new developments, it might be interesting to not only consider one of the two approaches, but also start investigating research designs that are both responsive and adaptive at the same time. Such designs could rely on different phases in which knowledge concerning the performance of design characteristics is gathered and then used in subsequent phases to optimize the survey design for specific subsamples. In earlier phases of the design, a researcher could test the application of different adaptive strategies. For example, if the focus is on improving participation across different strata of a sample, the effectiveness of group-specific treatments could be tested (e.g. younger respondents receive higher incentives, single person households are contacted more frequently, or older persons are offered different mode choices, etc.). Based on the evaluation of these adaptive strategies, later phases of a survey

could implement a design that includes the best-performing strategies and discards those that underperformed.

ACKNOWLEDGEMENT

Open access funding enabled and organized by ProjektDEAL.

ORCID

Tobias Gummer https://orcid.org/0000-0001-6469-7802

REFERENCES

AAPOR. (2016) Standard definitions: final dispositions of case codes and outcome rates for surveys. The American Association for Public Opinion Research.

Axinn, W.G., Link, C.F. & Groves, R.M. (2011) Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*, 48, 1127–1149. https://doi.org/10.1007/s13524-011-0044-1

Axinn, W.G., Wagner, J., Couper, M.P. & Crawford, S. (2021) Applying responsive survey design to small-scale surveys: campus surveys of sexual misconduct. *Sociological Methods & Research*, Online First, https://doi.org/10.1177/00491241211031270

Blom, A.G., Gathmann, C. & Krieger, U. (2015) Setting Up an online panel representative of the general population: the German internet panel. *Field Methods*, 27(4), 391–408.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A. et al. (2018) Establishing an open probability-based mixed-mode panel of the general population in Germany: the GESIS panel. *Social Science Computer Review*, 36(1), 103–115.

Church, A.H. (1993) Estimating the effect of incentives on mail survey response rates: a meta-analysis. *Public Opinion Quarterly*, 57(1), 62–79.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Daikeler, J., Bosnjak, M. & Lozar Manfreda, K. (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.

de Heij, V., Schouten, B. & Shlomo, N. (2015) RISQ Manual 2.1. Tools in SAS and R for the computation of R-indicators, partial R-indicators and partial coefficients of variation. http://www.risq-project.eu

de Leeuw, E., Hox, J. & Luiten, A. (2018) International nonresponse trends across countries and years: an analysis of 36 years of labour force survey data. *Survey Methods: Insights from the Field*.

Felderer, B., Muller, G., Kreuter, F. & Winter, J. (2018) The effect of differential incentives on attrition bias: evidence from the PASS wave 3 incentive experiment. *Field Methods*, 30(1), 56–69. https://doi.org/10.1177/1525822x17726206

Groves, R.M., Fowler, F.J., Couper, M.P., Singer, E. & Tourangeau, R. (2009) *Survey methodology*. Hoboken, NJ: Wiley.

Groves, R.M. & Heeringa, S.G. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society. Series A*, 169(3), 439–457.

Gummer, T. (2019) Assessing trends and decomposing change in nonresponse bias: the case of bias in cohort distributions. *Sociological Methods & Research*, 48(1), 92–115.

Hill, J., Smith, N., Wilson, D., White, J., & Richards, D. (2016). *Beginning postsecondary students longitudinal study (BPS: 12/14): data file documentation (NCES 2016-062)*. U.S. Department of Education, Washington, DC: National Center for Education Statistics.

Luijkx, R., Jónsdóttir, G.A., Gummer, T., Ernst Stähli, M., Frederiksen, M., Ketola, K. et al. (2021) The European values study 2017: On the way to the future using mixed-modes. *European Sociological Review*, 37(2), 330–346.

Mauz, E., von der Lippe, E., Allen, J., Schilling, R., Müters, S., Hoebel, J. et al. (2018) Mixing modes in a population-based interview survey: comparison of a sequential and a concurrent mixed-mode design for public health research. *Archives of Public Health*, 76(8), 1–17.

Murphy, K.R., Myors, B. & Wolach, A. (2014) *Statistical power analysis. a simple and general model for traditional and modern hypothesis tests*. New York, NY: Routledge.

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.

Pforr, K., Blohm, M., Blom, A.G., Erdel, B., Felderer, B., Fräßdorf, M. et al. (2015) Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly*, 79(3), 740–768.

Poushter, J. (2016) *Smartphone ownership and internet usage continues to climb in emerging economies*. Washington D.C.: Pew Research Center.

Raghunathan, T.E. & Grizzle, J.E. (1995) A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429), 54–63.

Schouten, B., Cobben, F. & Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101–113.

Schouten, B., Cobben, F., Lundquist, P. & Wagner, J. (2016) Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society A*, 179(3), 727–748.

Schouten, B., Peytchev, A. & Wagner, J. (2018) *Adaptive survey design*. Boca Raton, FL: CRC Press.

Singer, E. & Ye, C. (2013) The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 112–141. https://doi.org/10.1177/0002716212458082

Stoop, I., Billiet, J., Koch, A. & Fitzgerald, R. (2010) *Improving survey response. lessons learned from the european social survey*. Chichester: John Wiley & Sons.

Taylor, K. & Silver, L. (2019) *Smartphone ownership is growing rapidly around the world, but not always equally*. Washington D.C.: Pew Research Center.

Tourangeau, R., Brick, J.M., Lohr, S. & Li, J. (2017) Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society A*, 180(1), 203–223.

Villar, A. & Fitzgerald, R. (2017) Using mixed modes in survey research: evidence from six experiments in the ESS. In: Breen, M. (Ed.) *Values and identification. Evidence from the european social survey*. London, New York: Routledge, pp. 259–293.

Wagner, J. (2008) *Adaptive survey design to reduce nonresponse bias*. Ann Arbor, MI: University of Michigian. (PhD PhD thesis).

Weisberg, H.F. (2005) *The total survey error approach: a guide to the new science of survey research*. Chicago: University of Chicago Press.

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

How to cite this article: Gummer, T., Christmann, P., Verhoeven, S. & Wolf, C. (2022) Using a responsive survey design to innovate self-administered mixed-mode surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(3), 916–932. Available from: https://doi.org/10.1111/rssa.12835