Survey Research Methods (2025)
Vol. 19, No. 2, pp. 207-221
https://doi.org/10.18148/srm/2025.v19i2.8434
European Survey Research Association

Effects of Mode and Transitioning to a Mixed-Mode (Web/Phone) Design on Categorical Survey Estimates: Do Question Characteristics Matter?

Mengyao Hu[1,2] · Vicki A. Freedman[2] · Justin Kamens[3]
[1]School of Public Health, the University of Texas Health Science Center at Houston
[2]University of Michigan–Ann Arbor, Institute for Social Research
[3]Westat, Rockville, USA

An increasing number of telephone surveys are introducing a web option. Limited experimental research has explored the implications of such design changes and how they vary by question characteristics. Using an experimental design and propensity score matching techniques, this study examines the effects of mode (web vs. telephone) and transitioning to a mixed-mode design (from telephone-only to the choice of web or telephone) on survey estimates and how effects differ by question characteristics. We draw upon an experiment embedded in the National Study of Caregiving (NSOC), in which half of NSOC-eligible caregivers were randomized to a telephone-only design and the other half to a sequential mixed-mode design offering web and telephone options. For each categorical survey item, we test whether responses differ significantly by mode (after adjusting for selection effects using propensity score matching) and by design (experimentally assigned). We find that for the matched sample, significant differences by mode are evident for 10% of categorical survey items. These differences are larger for the approximately 75% of subjective (vs. 25% objective) questions and 45% non-binary (vs. 55% binary) questions, but are nevertheless negligible in size (<0.07 phi coefficient or Cramer's V). Because most effects are small and only about half of those randomized to the mixed-mode design opted for web, these mode effects rarely result in differences in estimates (3%) between the telephone-only and mixed-mode designs and differences in estimates are on average negligible in size (<0.03). We demonstrate that even if web take-up rates reach 90%, significant differences in estimates between telephone-only and mixed-mode designs remain rare (5%) and on average negligible in size (0.08) for the mix of categorical questions in NSOC. We discuss implications for designing future mixed-mode surveys.

Keywords: mixed-mode design; Question characteristics; Experimental design; Propensity score matching

1 Introduction

Mixed-mode data collection designs, in which web and interviewer-administered surveys are both offered as options for respondents, are becoming standard practice in survey research (Olson et al. 2021; Vannieuwenhuyze & Loosveldt 2013). Although mixed-mode surveys can increase response rates, minimize coverage error and reduce survey cost, they can also introduce mode effects. In the context of a panel study, changing to a mixed-mode design may potentially distort comparisons over time by introducing discontinuity in either item nonresponse or how questions are answered

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2025.v19i2.8434) contains supplementary material.

Corresponding author: Mengyao Hu, School of Public Health, the University of Texas Health Science Center at Houston, Houston, TX, USA (Email: mengyao.hu@uth.tmc.edu)

(Biemer et al. 2022; Cernat & Sakshaug 2020). The size of those distortions depends on both the percentage choosing to answer by web and the extent to which that choice influences responses (Vannieuwenhuyze & Loosveldt, 2013; Buelens & van der Brakel, 2015; Jäckle et al. 2010).

Investigations that attempt to uncover the influences of adding a web option to a survey have primarily focused on data quality indictors including unit and item nonresponse (Berete et al. 2019; Mackeben & Sakshaug 2023), length of response to open-ended questions (Chaudhary & Israel 2016) as well as field resources (Mackeben & Sakshaug 2023; McGonagle & Sastry 2023). Findings for these quality indicators have been mixed. For example, with respect to item nonresponse, Sastry and McGonagle (2022) found lower rates of missing data (less "skipping") for a representative sample of young adult respondents randomized to a sequential mixed-mode (web first with telephone follow-up) design in which 88% responded by web (vs. telephone-only design). In contrast, Ofstedal et al. (2022) found slightly higher item nonresponse among a representative sample of older adult respondents randomized to a sequential mixed-model design (vs. telephone-only design) in which the uptake of web was about 79%.

Research examining the impact of mixed-mode designs on survey estimates has focused on bias in specific domains or selected types of questions. For example, studies have examined non-differentiation of grid items with the same rating scales (e.g., Bowyer & Rogowski 2017), acquiescence for questions with agree-disagree scale (e.g., Cernat & Sakshaug 2020), sensitive content or questions subject to social desirability bias such as racial attitudes (e.g., Bowyer & Rogowski 2017; Cernat & Sakshaug 2020), fact-based knowledge questions where correct answers are known (e.g., Bowyer & Rogowski 2017), and demographic or employment-related items with analogues in high-quality administrative data (Sakshaug et al. 2023). On balance, these studies have concluded that there may be some differences in the use of rating scales and in answers to sensitive items, but measurement error bias is small for demographic and employment history items. The focus in previous studies on somewhat narrow question domains has yielded an incomplete picture of the extent to which the survey estimates are influenced by shifting to a mixed-mode design.

Other studies have relied on experimental designs that assign respondents directly to a single mode (e.g., web vs. telephone) in order to evaluate mode effects and how they vary by particular question characteristics. For example, Domingue et al. (2023) found evidence of higher cognitive scores for respondents randomly assigned to web-based data collection (vs. telephone); differences across modes were greatest for questions involving numbers (serial 7s and numeracy items). However, mode randomization experiments are not designed to provide insights into the extent to which shifting to a mixed-mode study could disrupt comparisons over time. Moreover, few studies have simultaneously considered multiple question features, limiting understanding of how question design choices might influence estimates.

This paper attempts to bridge existing lines of research by evaluating a randomized experiment implemented in the 2021 round of the National Study of Caregiving (NSOC). NSOC interviewed adult family-and-unpaid caregivers to participants in the National Health and Aging Trends Study (NHATS). The purpose of the experiment was to gauge the impact of changing from a telephone-only to mixed-mode (web-telephone) design on caregiving estimates. We explore for multiple question characteristics how survey estimates differ by: 1) mode (use of web vs. telephone), using a propensity score matching technique to address selection bias, and 2) design (mixed-mode vs. telephone-only), taking advantage of random assignment. We then illustrate the impact of web take-up rates in the mixed-mode study design on estimate differences between the telephone-only and mixed-mode designs.

2 Background and Hypotheses

2.1 Question Characteristics Influencing Responses by Mode

In a survey context, responses are the result of a process in which a respondent reacts to a question with particular characteristics. As shown in Fig. 1, in a mixed-mode context, the mode selected by the respondent (e.g. with an interviewer reading questions or by answering online) may influence how the task is perceived and hence how the question is answered.

At the individual level, mode choices can be influenced by respondent characteristics such as age, gender, education, and access to technology; such influences are sometimes referred to as "selection effects" (Vannieuwenhuyze & Loosveldt 2013). Controlling for selection effects, differences in estimates by mode (or "mode effects") may vary by question characteristics. Whether differences in estimates between mixed-mode and telephone-only designs (Box B in Fig. 1) are apparent depends in part on the size of the mode effects for different question types, the mix of questions in a study, and the percentage of respondents choosing to answer by web.

The literature provides guidance on how mode-related measurement effects vary by question characteristics. First, *subjective (attitude) questions* have been found to be more challenging than objective (factual) questions, especially if attitudes are not readily retrievable from memory (Bassili

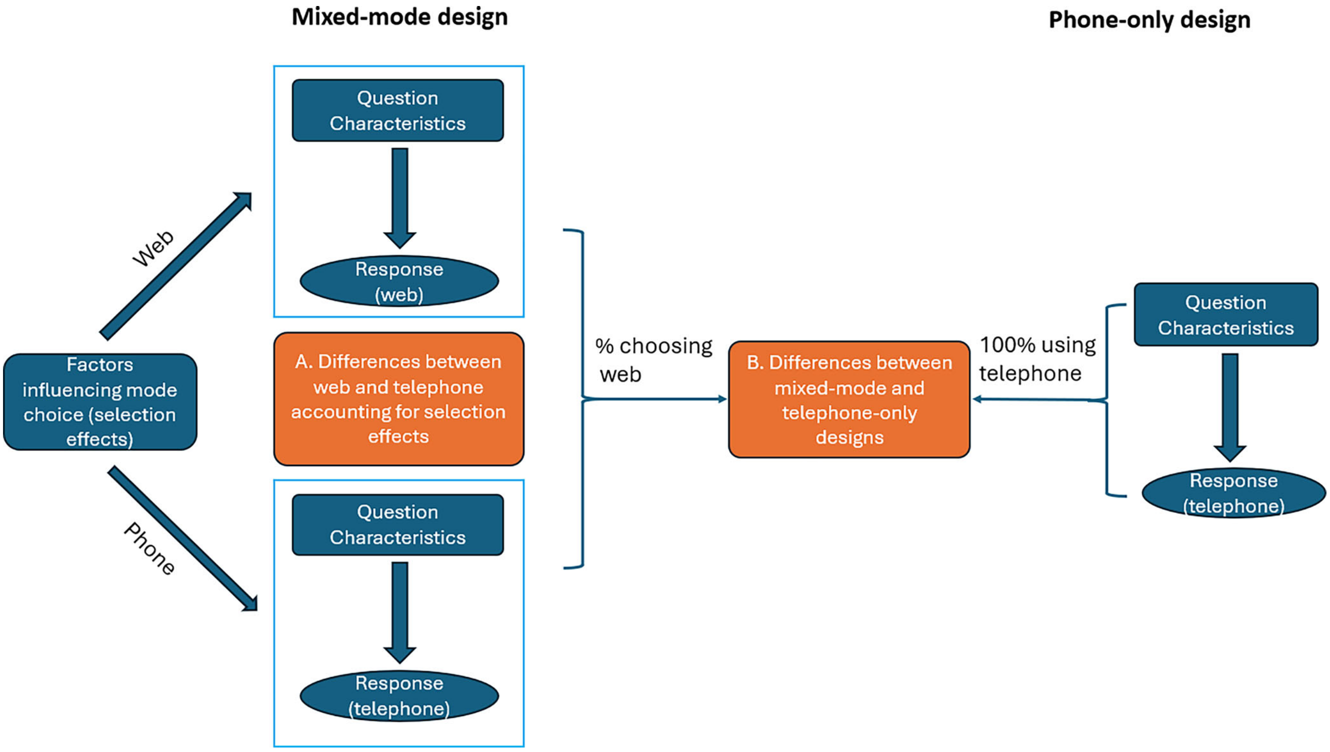

Fig. 1

Framework for Assessing Implications of Effects on Estimates When Changing to a Mixed-Mode (Web/Phone) Design

and Fletcher 1991; Tourangeau et al. 2000; Yan and Tourangeau 2008). This effect may be more likely in an interviewer-administered survey where there may be time pressure to provide an answer. Moreover, possibly due to social desirability bias, respondents tend to choose more positive response options to attitudinal questions in interviewer-administered modes relative to self-administered modes (Smyth, Christian, and Dillman 2008; Dillman et al. 2009). Second, for *nominal or unordered questions*, the number of outcome categories may have different effects by mode. For instance, a recency effect (i.e., choosing the last option) is more likely to occur in an interviewer-administered mode, whereas a primacy effect (i.e., choosing the first option) is expected when response options are presented visually (Stefkovics 2022). In contrast, for questions with ordered response scales, no clear response order effects are typically exhibited (Barlas & Thomas, 2012; Lynn et al. 2012; Stefkovics 2022). Third, although we are unable to address in this analysis, for completeness we note that *numeric questions* with continuous outcomes, often requiring estimation, pose greater difficulty for respondents than nominal, yes/no, and narrative questions in a telephone context (Olson et al. 2019). Hence, web surveys that allow respondents to set their own pace may result in fewer skips relative to telephone interviews (Chang and Krosnick 2009; Fricker et al. 2005) for such items.

2.2 Hypotheses

We draw upon the literature in formulating hypotheses by mode (web vs. telephone) and by design (telephone-only vs. mixed-mode) for different types of questions. We focus on two question characteristics that may be coded with relatively high accuracy (Bais et al. 2019) and that are prevalent in the NSOC instrument: (1) the type of content being asked (subjective/objective) and (2) the type and number of categories offered as answers (ordinal/nominal/binary; and number of categories).

H1 *Mode effects:* The differences in estimates between telephone and web modes will be:

a) More likely for subjective questions, relative to objective questions;
b) More likely for ordinal and nominal outcomes, relative to binary outcomes.
c) More likely for items with more response categories.

If the mode effects are large enough, the study has substantial numbers of items subject to those mode effects, and a substantial number of respondents in the mixed-mode group choose web, we are likely to also see the following effects when comparing the telephone-only and mixed-mode designs.

H2 *Design-related effects*: The differences in estimates between respondents randomized to a telephone-only and a mixed-mode (web/telephone) design will be:

a) More likely for subjective questions, relative to objective questions;
b) More likely for ordinal and nominal outcomes, relative to binary outcomes.
c) More likely for items with more response categories.

In addition, in sensitivity analyses, we test a counterfactual hypothesis as to whether having a choice of mode (rather than mode per se) yields differences in estimates (i.e., a "choice effect"). That is, we expect no differences in estimates between groups answering by telephone, when one group chose telephone, and the other was assigned to telephone. As a final step, we illustrate how differences in estimates by design (H2) change as the percentage of sample members opting to use web increases from 10 to 90%, given the mix of items currently in NSOC.

This study adds to the literature in several ways. First, using a matched sample to control for selection effects, we quantify effects on estimates by mode in a national survey context and examine how mode effects differ by several question characteristics. Second, drawing upon an experimental design, we quantify at the study level, the number of estimates influenced by a change in design from telephone-only to mixed-mode (web/telephone). Third, we illustrate how estimates vary by design as the percentage choosing web increases, holding the current mix of items constant.

3 Methods

3.1 Data

The National Study of Caregiving (NSOC) is a cross-sectional survey that interviews family and unpaid caregivers of participants in the National Health and Aging Trends Study (NHATS). The NHATS sample is drawn from the Medicare health insurance program, which covers about 96% of older adults in the United States. NHATS respondents were first interviewed in 2011 (71% response rate) and the sample was replenished in 2015 (77% response rate). The NHATS response rate was 94% in 2021, the focal year for this analysis (see Freedman et al. 2023a for additional details).

NSOC was designed to be an approximately 30-minute interview with family members and unpaid caregivers of NHATS participants who received help with self-care, mobility, or household activities. Up to five randomly selected caregivers for each NHATS participant are eligible to be interviewed in NSOC. The first three rounds of NSOC were conducted in 2011, 2015 and 2017 by telephone; response rates for cross-sectional samples ranged from 59.7 to 63%. The fourth round, conducted in 2021, had a stage 1 response rate (contact information provided) of 94% and a stage 2 response rate (caregiver participated, including partial and full responses) of 64%, yielding an overall response rate of 60%. Compared to all adults in the U.S. in 2021, (U.S. Census Bureau, 2023), the population of family caregivers to older Medicare beneficiaries is older (mean age 62 vs. 48 nationally); more likely to be female (65% vs. 51%), more likely to be non-Hispanic White (70% vs. 62%) and less likely to be Hispanic or Latino (7% vs. 17%).

The 2021 round of NSOC included an experiment, in which eligible caregivers (n = 3216) were randomly assigned to a mixed-mode (web or telephone) design (n = 1600 respondents) or to a telephone-only design (1616 respondents). All caregivers for a given NHATS participant were randomized together (that is, to the same design group). The purpose of the experiment was to assess the impact of changing from a telephone to a mixed mode (web/telephone) design.

3.2 Experimental Protocols

3.2.1 Initial Invitation

The NHATS participant was asked to provide contact information for each eligible NSOC caregiver and to pass along a packet of information about the study. Eligible caregivers were then contacted and invited to participate in the study. In the group randomized to the mixed-mode design, recruitment protocols varied by available contact information. For those with an available mailing address (81%), a welcome letter with web invitation, information sheet, and $20 prepaid incentive check were sent by mail. Two days later, those with an available email address were sent an email version of the invitation. For respondents without a mailing address (19%), interviewers called to invite them to participate in NSOC, offered to send a $20 incentive check, and gave the respondent a choice of doing the interview by telephone or web.

For the telephone-only design group, if address information was available (76%), a welcome letter, information sheet, and $20 prepaid incentive check were sent by mail. Interviewers called respondents seven days later to conduct the telephone interview. If address information was not available (24%), interviewers called the respondent on the day the case was released, invited them to participate in the interview by telephone, and offered to send a $20 incentive check.

3.2.2 Follow-up Reminders

For both design groups, depending on contact information availability, reminder letters, emails and texts were sent to respondents every seven days until 4–5 weeks after the case was released. For mixed-mode respondents only, each reminder included login information and a direct link to log in to the survey. For both groups, with the fourth and final reminder, an additional $20 postpaid incentive was offered to those who had not yet responded.

Reminder calls were made after 14 or 7 days to respondents with and without an email address, respectively. During the reminder call, interviewers also offered respondents the telephone interview option. Telephone call reminders continued until a maximum number of calls (7) was reached. An additional set of telephone calls took place for those who had not yet responded during weeks 7 or 8.

3.2.3 Response Rates, Item Nonresponse and Field Effort

The response rate (AAPOR's RR2) was 62% for the mixed-mode design group and 58% for the telephone-only group. Item nonresponse was generally low and did not differ by mixed-mode (3% on average) vs. telephone-only (2%) designs. The mixed-mode design entailed fewer outbound telephone calls per respondent than the telephone-only design (4.25 vs. 5.00), but other contacts (e.g., emails, mails, and texts) did not differ (Freedman et al., 2023b).

3.3 Items

The NSOC questionnaire covers a variety of topics about the caregiving experience. Questionnaire sections focused on care activities carried out for the NHATS participant, duration of care, interactions with health care providers, positive and negative aspects of caregiving, the support environment, visits with the NHATS participant, and distance to the NHATS participant. In addition, sections of the questionnaire ask about other aspects of the caregiver's life including participation in valued activities, their health (including experiences with COVID-19), household and demographic factors, race/ethnicity, employment and caregiving, and health insurance and income. Minor changes were made to the telephone questionnaire to transform it into a web instrument. For example, over the telephone, interviewers read the response options as part of the question; on the web, the response options were visible to the respondent, but not as part of the question. For a small number of items (e.g., duration of help, hours of work missed), the telephone instrument allowed respondents to answer in different units (e.g. number of years or the year started helping) whereas the web respondents were not given a choice (How many years have you ...). For questions about dollar amounts (spent, given, received), the telephone instrument offered a few categories in an initial question and higher or lower categories in a follow-up question whereas the web instrument displayed all categories at once. In addition, in telephone interviews, interviewers recorded item nonresponse as "Don't know" or "Refused" when respondents volunteered these answers; in contrast, in the web mode, respondents were not offered "Don't know" or "Refused" answers, but they were able to skip a question without answering by advancing to the next screen. Finally, for a few items, some responses (e.g., whether respondent is retired or doesn't work anymore) were not read but recorded if volunteered in the telephone interview but presented to all respondents in the web mode.

Altogether 262 items were reviewed for inclusion. We first removed 30 continuous items (so that we could focus on variation by response scale type (binary, ordinal, nominal) and number of response options for all items), thus limiting analyses to 232 categorical items. We further excluded 6 demographic items because such information is readily accessible and therefore less subject to measurement error (Olson et al. 2019; Sakshaug et al. 2023). We also excluded 3 open-ended questions because their responses require manual coding, which may introduce additional sources of bias. The final sample of items included 223 categorical items; the full list can be found in Hu and Freedman (2023).

For each item, question type was coded by two coders independently. Inter-coder reliability was 0.72, suggesting good agreement between the two coders. Questions with inconsistent coding were reviewed and recoded by two expert survey methodologists. For response scale types, we combined nominal and select-all-that-apply question due to a small number of question items fall into these categories. Across the 223 items, 26% were classified as objective and 74% as subjective; 55% were binary, 37% ordinal and 8% nominal; and 11% had 3, 16% had 4, and 18% had 5 or more response options.

3.4 Respondent Samples

Of the 1938 NSOC participants, 942 were assigned to the telephone-only and 996 to the mixed-mode design. Among mixed-mode design respondents, 555 initially responded by web (56%) and the remaining 441 (44%) initially responded by telephone. Fewer than 20 cases switched modes (typically later in the interview), so they were classified for this analysis by their initial mode.

3.4.1 Analytical Sample for Web vs. Telephone (H1)

For mode effect hypotheses (H1), we performed propensity score matching to create a comparison sample for those who chose web. The matched sample was selected among those who were assigned to the telephone-only design but who had a high propensity for choosing web (had they been given a choice). We first estimated propensity scores using a weighted logistic regression to predict mode choices, with covariates in Appendix Table A2. We included a set of demographic variables (e.g. age, relationship to care recipient, gender, education level, marital status, whether the caregiver has children under age 18, race/ethnicity, metro/non-metro residence, co-residence with the care recipient, household size) and type of contact information available (address, phone, email). We chose these indicators because they may be related to mode choice and to study outcomes of interest (e.g., Brookhart et al. 2006; Chen et al. 2022; Kibuchi et al. 2024). We then matched 1:1 telephone-only respondents to those who chose web using nearest neighbor within caliper matching method. A caliper size of a quarter of a standard deviation of the sample estimated propensity scores was used (Rosenbaum & Rubin 1985; Guo et al. 2020). This approach allows us to disentangle differences in sample composition between respondents who answered using different survey modes by ensuring similar background characteristics across groups (Lugtig et al., 2011). That is, differences observed after matching are likely due to mode effects rather than confounding selection effects. This approach yielded 528 respondents (out of 555 respondents whose initial mode was web) matched to 528 respondents who were assigned to the telephone-only design. Analysis for H1 are restricted to these 1056 respondents. To examine the effectiveness of the propensity score in adjusting for confounding factors, we evaluated matching quality using two methods: 1) comparisons of standardized mean differences (SMDs) before and after matching (Austin, 2009; Kibuchi et al. 2024); 2) comparisons of covariates distributions before and after matching using chi-square test for binary or categorical variables and t-test for continuous variables. For SMDs, covariate balance in matched samples is typically considered achieved when all SMDs are below 0.10 (Austin 2011). However, because the 0.10 threshold is somewhat arbitrary, moderate imbalances of SMD <0.25 can also be considered acceptable, especially in small samples (e.g., Austin, 2009). These evaluations, shown in Appendix Tables A1, confirm that the matching resulted in balanced comparisons groups on all factors considered.

3.4.2 Analytical Sample for Telephone with a Choice vs. Telephone Without Choice (Sensitivity Analysis)

We also created an additional matched sample consisting of individuals in the mixed-mode design who chose to answer by telephone with those assigned to the telephone-only design who had a high propensity for choosing telephone (see Appendix Table A1) using the same matching method and evaluation approach describe earlier. This matched sample was created to test the counterfactual argument that having a choice in and of itself (rather than which mode is selected) is not influencing outcomes and, when used in combination with the matched sample of those choosing web to those with a propensity for choosing web, to allow us to illustrate the impact on estimates of the extent of web take-up in the mixed mode design. In total, 418 respondents (out of 441 respondents who initially chose the telephone) were matched to 418 respondents who were assigned to telephone-only design. After matching, all variables except race/ethnicity were no longer significantly different between the two groups, and SMD for race/ethnicity was reduced from 0.31 before matching to 0.24 after matching.

3.4.3 Analytical Sample for Mixed-mode vs. Telephone-only (H2)

For design-related effect hypotheses (H2), we compared responses for individuals randomized to the mixed-mode design (n = 996) to those randomized to telephone-only design (n = 942). Before undertaking comparisons of outcomes of interest, we explored the balance between the two groups with respect to the caregivers' gender, relationship to the NHATS participant, types of contact information available (e.g., address, telephone number, email), and variables reflecting the NHATS participant's demographic and socioeconomic characteristics. We confirmed that the randomization yielded substantially balanced design groups (see Appendix Table A3).

3.4.4 Analytical Sample for Varying Percentage of Respondents Choosing Web

To illustrate how design-related effects change with the percentage of respondents choosing web, we drew upon the two matched samples previously described. The first matched sample includes (A) mixed-mode design respondents choosing web (n = 528) matched to (B) telephone-only design respondents with a high propensity to choose web (n = 528). The second matched sample is from the previously mentioned sensitivity analysis, which includes (C) mixed-mode design respondents choosing telephone (n = 418) matched to (D) telephone-only design respondents with a high propensity to choose telephone (n = 418)[1]. The final mixed-mode sample is constructed with Groups A and C; and the final telephone-only sample includes Groups B and D. This approach results in a matched sample between the mixed-mode and telephone-only designs, referred to as the "matched mixed-mode design sample." This matched sample serves as the basis for our analysis examining effects of varying percentage of respondents choosing web (details described below).

3.5 Measures

3.5.1 Outcomes

For each categorical item, we examined whether estimates differed by mode (using matched samples) and by design (using randomized samples) using weighted chi-square tests. Given that a large number of tests were conducted, the Benjamini-Hochberg (BH) method was used to adjust significance of the tests to reduce the false discovery rate (hereafter "adjusted tests"; see Thissen et al. 2002). For each item, we then generated a binary indicator indicating whether estimates differed significantly—by mode (for H1) and by design (for H2). In addition, we calculated an effect size for each difference using either a phi coefficient (for binary outcomes) or a Cramer's V statistic (for non-binary outcomes). By convention, effect sizes <0.09 indicating a negligible effect, 0.10–0.29 a small effect and >0.29 a medium or large effect.

3.5.2 Predictors

We focus on question type (subjective vs. objective), response scale type (binary, ordinal, and nominal), and number of response options (excluding don't know or refused). We also controlled for item-level sample size, which can vary depending on skip patterns.

3.6 Analytic Approach

We estimated logistic regression models[2] to predict the significance of adjusted chi-square tests summarizing differences in estimates by mode (H1) and by design (H2). Coefficients can be interpreted as the difference in the log-odds of having a significant mode/design effect for a given question characteristic. We also estimated beta regression models (Geissinger et al. 2022) to predict effect sizes for differences in estimates by mode (H1) and by design (H2). Coefficients indicate whether a given item characteristic is associated with higher or lower effect size (relative to the omitted category). To facilitate interpretation, we also report average marginal effects for covariates of interest. These effects show the magnitude of average increase or decrease of effect size for a specific category in relation to the reference category for a one-unit increase in the covariate. Covariates include question type (subjective vs. objective), response scale type (binary, ordinal, and nominal), and number of response options (as a continuous variable). Because significance is in part a function of sample size, and the number of individuals asked a given item varied by skip and related nonresponse patterns, we also controlled for the number of respondents who were asked the given item. With the unit of analysis being at the variable level, both models are unweighted.

Finally, we illustrate how the percentage of estimates that differ between telephone-only and mixed-mode designs changes as the percentage of respondents choosing web in the mixed-mode design increases using the matched mixed-mode design sample described above. We used a descriptive weighting adjustment approach, which involves applying varying adjustment factors to modify the total sum of NSOC survey weights for each of the two mode groups in the mixed-mode design. The rationale for this approach is to control the relative composition of each group in the mixed-mode design. We examined nine scenarios with the percentage of web respondents ranging from 10 to 90% and the corresponding percentage of telephone-only respon-

[1] Note that 204 respondents appear in both B and D. Given this is a relatively small percentage (12% of 1,688 unique respondents in groups A–D) and the goal to have balanced sample between mixed-mode vs. telephone-only conditions, we treated them as if they were different respondents in our analysis.

[2] Note that we are unable to directly model estimates for all items in the analysis using a multilevel model, given that outcomes differ by item. Instead, we model whether the difference for each item is significant or not by mode and design using logistic regression.

Table 1

Percentage of items with significant differences and effect size by mode (web vs. telephone). (Source: National Study of Caregiving)

		Estimates	
	%	% items with significant difference by mode[a]	Effect size
Overall	–	10	0.08
Question types		*	**
Objective	27	7	0.07
Subjective	74	19	0.10
Response scale types		***	***
Binary	55	3	0.04
Ordinal	37	18	0.12
Nominal & others	9	21	0.14
Number of response options		***	***
2 (Binary)	56	3	0.04
3	11	8	0.12
4	16	33	0.14
5+	17	13	0.13

n = 223 categorical items answered by *n* = 1056 matched respondents. Note that respondent sample sizes differ by item (as a result of varying skip patterns and item nonresponse)

[a] Based on Benjamini-Hochberg-adjusted chi-square tests

* p < 0.05; ** p < 0.01; *** p < 0.001

dents from 90 to 10%. The total sum of weights was fixed across all nine scenarios. In the first scenario, we adjusted the sum of weights for those who choose web to be 10% of the total sum of weights, and the sum of weights for those who choose telephone was adjusted to 90% of the total sum of weights. This adjustment was accomplished by multiplying each web respondent's weight by

$$\frac{(0.1*\text{total sum of weights})}{\begin{array}{c}\text{sum of weights for those who}\\ \text{choose web before adjustment}\end{array}},$$

and similarly, multiplying each phone respondent's weight in the mixed-mode group by

$$\frac{(0.1*\text{total sum of weights})}{\begin{array}{c}\text{sum of weights for those who}\\ \text{choose phone before adjustment}\end{array}},$$

We follow the same adjusting logic for the remaining eight scenarios. Finally, using each of the nine sets of weights, we computed descriptive measures summarizing results for all items, including percentages with significant differences in estimates between the mixed-mode vs. telephone-only designs (using BH adjusted Chi-square tests) and mean effect sizes (using phi coefficients or Cramer's V statistics).

4 Results

4.1 Estimates by mode (web vs. Telephone) in Matched Samples

Estimates differ by mode for 10% of items (Table 1). Consistent with H1a, b and c, mode effects were more likely to be observed for subjective questions (vs. objective questions) and differed by response scale type (more likely for ordinal and nominal) and by number of response options (most likely for 4 categories). Mean effect sizes also differ by question characteristics, but are negligible or small for all categories (0.04–0.14). The largest significant differences by mode were evident for three employment questions about work for pay (effect size 0.28–0.42, considered moderate sized) and notable differences also appeared for the last five items in a seven-item series measuring caregiver wellbeing (effect size 0.15–0.23, considered small) using a four-category scale from "agree strongly" to "disagree strongly."

Models predicting differences in estimates by mode that control for question type, response scale type and number of response options (Table 2) suggest that response scale type is the salient question characteristic predicting mode effects (H1b). Results from the beta regression model suggest that effect size is higher for ordinal and nominal questions than

Table 2

Models predicting significant differences in estimates and effect sizes by mode (web vs. telephone). (Source: National Study of Caregiving)

	Logistic regression predicting significant differences in estimates across modes		Beta regression predicting effect size differences across modes		
	Coefficient	Std. Err.	Coefficient	Std. Err.	Average marginal effects
Question types					
Objective (reference)					
Subjective	0.17	0.64	0.10	0.11	0.01
Response scale types					
Binary (reference)					
Ordinal	2.16*	0.87	0.91***	0.13	0.07
Nominal & others	2.75**	0.87	0.75***	0.15	0.05
Number of response options	–0.10	0.18	0.05*	0.02	0.00
Respondent sample size[a]	0.001	0.001	–0.001***	0.000	–0.000

n = 223 categorical items answered by n = 1056 matched respondents. Note that respondent sample sizes differ by item (as a result of varying skip patterns and item nonresponse)

[a]The mean respondent sample size is n = 759

* p < 0.05; ** p < 0.01; *** p < 0.001

binary questions and increases with the number of response options; effect sizes also decline with respondent sample size. Average marginal calculations indicate that both ordinal and nominal/other variables increase effect size by 0.07 and 0.05, respectively, in comparisons to binary variables, which are negligible differences in size.

4.2 Estimates by Design

Estimates differ by design for only 3% of items (Table 3). The largest significant difference by design was evident for a question about making medical decisions in the last month of the care recipient's life (effect size 0.23, considered small). We do not find support for H2a, b and c; that is, effects by design do not vary by question type, response scale type, or consistently by the number of response options. We do observe that questions with 4 response options stand out as having a qualitatively higher percentage of items with a significant difference by design group (11%) compared with other category counts (0.0–4%). Mean effect sizes, although they vary by response scale type and number of response options, are negligible or small for all categories examined (≤ 0.1).

Models predicting differences in estimates by design suggest effects do not vary by the three question characteristics that we investigated (Table 4; H2b). Results from the beta regression model predicting effect size suggest that ordinal variables, variables with more response options and smaller test sample size tend to have larger effect sizes; however, marginal calculations indicate that differences of the effect sizes by the type of content being asked and the characteristics of answers are on average negligible in size (< 0.03).

4.3 Sensitivity: Estimates by Whether the Respondent has a Choice to Respond by Telephone

As expected, that there are no effects on estimates of choosing (vs. random assignment to) telephone, in descriptive analyses (see Appendix Table A4); we therefore refrained from estimating models.

4.4 Summary of Hypotheses

Table 5 provides a summary of findings, including whether there is descriptive and model-based support for each hypothesis by mode and design and, if so, the size of the effect. For descriptive analyses, differences by mode are rare and negligible or small in size and there is no evidence of significant differences by design. For model-based analysis of differences by mode, the likelihood of having significantly different estimates based on nominal and ordinal questions is greater than for binary questions but differences are negligible in size. For model-based analysis of differences by design, the likelihood of having significantly different estimates does not vary by question characteristics.

Table 3

Percentage of items with significant differences in estimates and effect size by design (mixed-mode vs. telephone-only). (Source: National Study of Caregiving)

		Estimates	
	%	%^{a}	Effect size
Overall	100	3	0.07
Question types			
Objective	27	7	0.07
Subjective	74	2	0.06
Response scale types			***
Binary	55	2	0.04
Ordinal	37	5	0.09
Nominal	9	5	0.08
Number of response options		*	***
2 (Binary)	56	2	0.04
3	11	4	0.08
4	16	11	0.10
5+	17	0	0.10

n = 223 categorical items answered by *n* = 1938 randomly assigned respondents. Note that respondent sample sizes differ by item (as a result of varying skip patterns and item nonresponse)
[a]Based on Benjamini-Hochberg-adjusted chi-square tests
* p <0.05; ** p <0.01; *** p <0.001

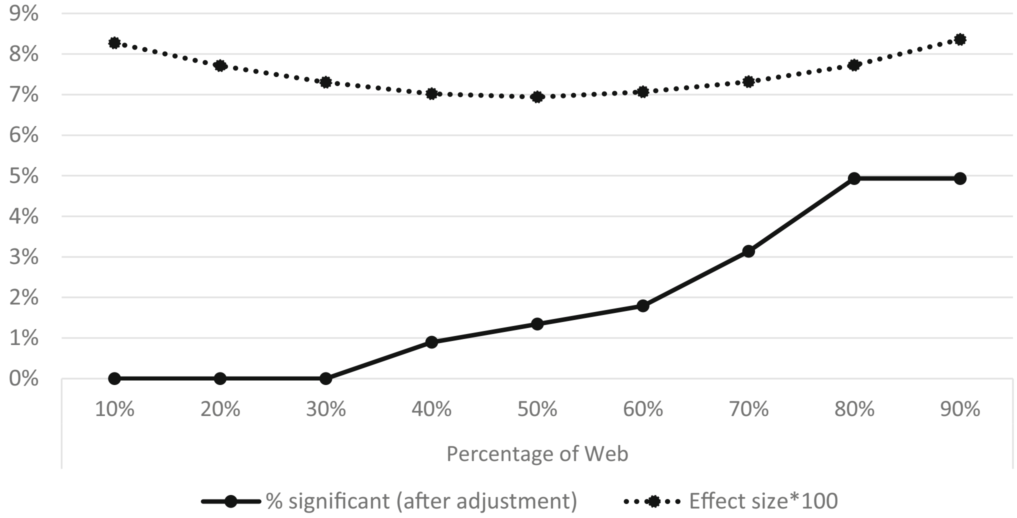

Fig. 2

*Percentage of items with significant differences between mixed-mode and telephone-only designs (after BH adjustment) and effect size (*100) for 223 items in NSOC, by percentage choosing web in the mixed-mode design*

Table 4

Models predicting significant differences in estimates and effect sizes by design (mixed-mode vs. telephone-only). (Source: National Study of Caregiving)

	Logistic regression predicting significant differences in estimates across designs		Beta regression predicting effect size differences across designs		
	Coefficient	Std. Err.	Coefficient	Std. Err.	Average marginal effects
Question types (ref. categ.: objective)					
Objective (reference)					
Subjective	1.19	1.14	0.09	0.12	0.01
Response scale types (ref. categ.: binary)					
Binary (reference)					
Ordinal	1.06	1.54	0.52***	0.14	0.03
Nominal & Others	2.03	1.49	0.22	0.17	0.01
Number of response options	–0.27	0.50	0.09***	0.03	0.01
Respondent sample size[a]	0.000	0.000	–0.001***	0.000	–0.000

n = 223 categorical items answered by n = 1938 matched respondents. Note that respondent sample sizes differ by item (as a result of varying skip patterns and item nonresponse)

[a] The mean respondent sample size is n = 1363

* p <0.05; ** p <0.01; *** p <0.001

4.5 Illustration of Differences in Estimates by Design by Web Take-up Rates

Using the mixture of items in the NSOC interview, Fig. 2 illustrates the percentage of items that would yield significantly different estimates and mean effect sizes (*100) by design, as the percentage answering by web increases from 10 to 90% of the sample. As expected, the percentage of estimates that differ between the mixed-mode and telephone-only designs increases as more respondents in the mixed-mode choose to respond by web. In this illustration, the estimate of differing items reaches a plateau at about 5% of variables yielding different estimates when 80% of respondents choose web. Effect sizes are on average negligible in size (0.07–0.08 range) irrespective of the percentage choosing web.

5 Discussion

This study examined differences in survey estimates by mode (phone vs. web) and by study design (phone-only vs. mixed-mode) for more than 200 categorical items in a national survey of caregivers. Several findings are noteworthy. First, in descriptive analyses of a matched sample comparing estimates by telephone and web, about one in ten items produced significantly different estimates; yet these differences were not appreciable in size. Second, in models controlling for multiple item characteristics, nominal and ordinal questions were more susceptible than binary measures to mode effects, but differences were small. Third, given the relatively small effects and the fact that only about half of the mixed-mode sample opted to complete the study by web, only 3% of items yielded different estimates between telephone-only and mixed-mode designs; this finding did not differ appreciably by question characteristics that we investigated. Finally, we illustrated that even if web take-up rates in the mixed-mode design reached 90%, items with significant differences in estimates between the telephone-only and mixed-mode designs remain rare (5%) and on average small for the mix of items in NSOC.

Our findings suggest that questionnaire design decisions can play a role in driving mode effects. Analysis revealed that three employment-related questions significantly differed between web and telephone modes. These moderately-sized differences likely stem from offering a “retired/doesn’t work anymore” response option to web respondents, but recording this response in telephone interviews only when volunteered by respondents. When this category is collapsed with an alternative response, such as “No” for the whether-work-for-pay question, these differences are no longer observed. Additionally, responses to several well-being questions (using options “Agree strongly,” “Agree somewhat,” “Disagree somewhat,” and “Disagree strongly”) exhibited small mode differences, with telephone respondents more inclined toward extreme responses showing positive well-being compared to web respondents. This difference may result from social desirability bias in telephone interviews, given the sensitive nature of these questions, or more likely, due to differences in response option presenta-

Table 5

Summary of Findings for H1 and H2

Hypothesis	Descriptive Support for Significant Differences by:		Model-based Support for Significant Differences by:	
Difference in Estimates is	Mode	Design	Mode	Design
Greater for subjective questions, relative to objective questions	H1a Yes (negligible or small)	H2a No	H1a No	H1a No
Greater for nominal and ordinal outcomes, relative to binary outcomes	H1b Yes (negligible or small)	H2b No	H1b Yes (negligible or small)	H1b No
Greater for items with more response categories	H1c Yes (negligible or small)	H2c No	H1c No	H1c No

tion. Specifically, web respondents see all options at once, while telephone respondents hear the options only once at the beginning of this series, resulting in mode effects that particularly impact on the last items in the series. Beginning in NSOC Round 13, response options in the telephone mode were repeated for the first, second, fourth and fifth items. Notably, when response options are dichotomized to "agree" vs. "disagree," the significant mode differences for these items are no longer observed.

This study is not without limitations. Our focus was exclusively on categorical items and does not provide insight into continuous (numeric) responses. However, others have demonstrated the greater difficulty answering numerical items than other types of questions in a telephone context (Olson et al., 2019), and differences in estimates between mixed-mode and telephone-only designs for numeric items along a scale reflecting expectations (Ofstedal et al. 2022). Such differences may be linked to heightened cognitive effort related to numeric responses (Lipps and Monsch, 2022), although more research is needed on this point. We also did not focus on constructs that involve multiple measures (Sakshaug et al. 2022), or on the direction of estimate differences, or on the role of response style (Van Vaerenbergh & Thomas 2013). The latter may be particularly important for future research in light of findings that acquiescence and extreme responses are more common in interviewer-administered surveys than in web surveys (Liu et al. 2017; Weijters, Geuens, & Schillewaert 2008; Vaerenbergh & Thomas 2013). In addition, matching among respondents may not fully eliminate all selection bias, as it is possible that unobserved variables might be absent from the set of matching variables in the propensity score models. Given the range of outcomes analyzed in this paper, it is also possible that the effectiveness of matching in reducing selection bias may vary across different variables, being more successful for some and less so for others. Finally, we did not directly examine the implications of shifting to a mixed-mode design on trend estimates, an important next step.

Despite these limitations, this study has both theoretic and practical implications for researchers working in a mixed-mode survey context. Theoretically, we do find some differences in response patterns by mode that are more likely to emerge when questions are subjective, nominal, and ordinal and with more categories (relative to binary questions), yet these differences are negligible or small so that even at high levels of web uptake, a change in design from telephone-only to mixed-mode does not result in appreciably different estimates for the vast majority of categorical items.

Results of this study also have practical implications. Findings suggest that survey analysts working with NSOC can confidently combine data from both telephone and web modes. In addition, we demonstrated that even very high levels of web take-up by those offered a mixed-mode design are unlikely to introduce meaningful discontinuities for most estimates. These findings are dependent on the mix of questions in NSOC (26% objective; 55% binary); it would be worthwhile to explore how question mix might influence these findings either illustratively in NSOC or with other studies that have a different mix of questions. Nevertheless, combined with the lack of appreciable differences by question type in this study, findings are sufficiently encouraging that NSOC's switch from telephone-only to a mixed-mode design is unlikely to perturb longer-term trend analysis. Finally, the study offers guidance to survey researchers designing mixed-mode studies. Since subjective questions and those with more response categories are more susceptible to mode effects (even if small), findings suggest that when in doubt a binary, objective item has the best chances of minimizing mode and design effects.

Acknowledgements We thank Mick Couper for helpful comments on earlier drafts and for his contributions to the design and implementation of the mixed-mode experiment embedded in the National Study of Caregiving.

Funding This research was funded by the National Institute on Aging of the National Institutes of Health (U01AG032947 and R01AG062477).

References

Austin, P.C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, *28*(25), 3083–3107.

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399–424.

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tòth, J., Douhou, S., & Vis, C. (2019). Can survey item characteristics relevant to measurement error be coded reliably? A case study on 11 Dutch general population surveys. *Sociological Methods & Research*, *48*(2), 263–295.

Barlas, F.M., & Thomas, R.K. (2012). "Economic confidence: Effects of response format in trend sensitivity and correspondence with national measures," In 67th annual conference of the American Association for Public Opinion Research, Orlando, FL. http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2012/05_BarlasThomas_H4_FINAL.pdf.

Bassili, J.N., & Fletcher, J.F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, *55*(3), 331–346.

Berete, F., Van der Heyden, J., Demarest, S., Charafeddine, R., Gisle, L., Braekman, E., & Molenberghs, G. (2019). Determinants of unit nonresponse in multimode data collection: a multilevel analysis. *Plos one*, *14*(4), e215652.

Biemer, P.P., Harris, K.M., Burke, B.J., Liao, D., & Halpern, C.T. (2022). Transitioning a panel survey from in-person to predominantly web data collection: Results and lessons learned. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 798–821.

Bowyer, B.T., & Rogowski, J.C. (2017). Mode matters: evaluating response comparability in a mixed-mode survey. *Political Science Research and Methods*, *5*(2), 295–313.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, *163*(12), 1149–1156.

Buelens, B., & van den Brakel, J.A. (2015). Measurement error calibration in mixed-mode sample surveys. *Sociological Methods & Research*, *44*(3), 391–426.

Cernat, A., & Sakshaug, J. (2020). The impact of mixed-modes on multiple types of measurement error. *Survey Research Methods*, *14*(1), 79–91.

Chang, L., & Krosnick, J.A. (2009). National surveys via RDD telephone interviewing versus the Internet: comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678.

Chaudhary, A.K., & Israel, G.D. (2016). Assessing the influence of importance prompt and box size on response to open-ended questions in mixed-mode surveys: Evidence on response rate and response quality. *Journal of Rural Social Sciences*, *31*(3), 7.

Chen, J.W., Maldonado, D.R., Kowalski, B.L., Miecznikowski, K.B., Kyin, C., Gornbein, J.A., & Domb, B.G. (2022). Best practice guidelines for propensity score methods in medical research: consideration on theory, implementation, and reporting. A review. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, *38*(2), 632–642.

Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B.L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social science research*, *38*(1), 1–18.

Domingue, B., McCammon, R., West, B., Langa, K., Weir, D., & Faul, J. (2023). The mode effect of web-based surveying on the 2018 U.S. Health and retirement study measure of cognitive functioning. *The Journals of Gerontology: Series B*, *2023*, gbad68. https://doi.org/10.1093/geronb/gbad068.

Freedman, V.A., Schrack, J.A., & Skehan, M.E. (2023a). *National health and aging trends study user guide: rounds 1–12 final release*. Baltimore: Johns Hopkins Bloomberg School of Public Health.

Freedman, V.A., Hu, M., & Wolff, J. (2023b). *National study of caregiving IV user guide: : rounds 11–12 final release*. Baltimore: Johns Hopkins Bloomberg School of Public Health.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Geissinger, E.A., Khoo, C.L., Richmond, I.C., Faulkner, S.J., & Schneider, D.C. (2022). A case for beta regression in the natural sciences. *Ecosphere*, *13*(2), e3940.

Guo, S., Fraser, M., & Chen, Q. (2020). Propensity score analysis: recent debate and discussion. *Journal of the Society for Social Work and Research*, *11*(3), 463–482.

Hu, M., & Freedman, V. A. (2023). *Effects of switching to mixed-mode design on estimates from NSOC IV 2021 (round 11)*. NHATS Technical Paper #40. Baltimore: Johns Hopkins Bloomberg School of Public Health.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3–20.

Kibuchi, E., Sturgis, P., Durrant, G. B., & Maslovskaya, O. (2024). The efficacy of propensity score matching for separating selection and measurement effects across different survey modes. *Journal of Survey Statistics and Methodology*, *12*(3), 764–789.

Lipps, O., & Monsch, G. A. (2022). Effects of question characteristics on item Nonresponse in telephone and web survey modes. *Field Methods*, *34*(4), 318–333.

Liu, M., Conrad, F. G., & Lee, S. (2017). Comparing acquiescent and extreme response styles in face-to-face and web surveys. *Quality & quantity*, *51*, 941–958.

Lugtig, P., Lensvelt-Mulders, G. J., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669–686.

Lynn, P., Hope, S., Jäckle, A., Campanelli, P., & Nicolaas, G. (2012). *Effects of visual and aural communication of categorical response options on answers to survey questions*. (No. 2012–21). ISER Working Paper Series.

Mackeben, J., & Sakshaug, J. W. (2023). Transitioning an employee panel survey from telephone to online and mixed-mode data collection. *Statistical Journal of the IAOS*, *39*(1), 213–232.

McGonagle, K. A., & Sastry, N. (2023). Transitioning to a mixed-mode study design in a national household panel study: effects on fieldwork outcomes, sample composition and costs. *Survey research methods*, *17*(4), 411.

Ofstedal, M. B., Kézdi, G., & Couper, M. P. (2022). Data quality and response distributions in a mixed-mode survey. *Longitudinal and Life Course Studies*, *13*(4), 621–646.

Olson, K., Smyth, J. D., & Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, *7*(2), 275–308.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., & Wagner, J. (2021). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*, *9*(3), 381–411.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903.

Sakshaug, J. W., Cernat, A., Silverwood, R. J., Calderwood, L., & Ploubidis, G. B. (2022). Measurement equivalence in sequential mixed-mode surveys. *Survey Research Methods*, *16*(1), 29–43.

Sakshaug, J. W., Beste, J., & Trappmann, M. (2023). Effects of mixing modes on nonresponse and measurement error in an economic panel survey. *Journal for Labour Market Research*, *57*(1), 2.

Sastry, N., & McGonagle, K. A. (2022). Switching from telephone to web-first mixed-mode data collection: results from the transition into adulthood supplement to the US panel study of income dynamics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *185*(3), 933–954.

Smyth, J. D., Christian, L. M., & Dillman, D. A. (2008). Does “yes or no” on the telephone mean the same as “check-all-that-apply” on the web? *Public Opinion Quarterly*, *72*(1), 103–113.

Stefkovics, Á. (2022). Are scale direction effects the same in different survey modes? Comparison of a face-to-face, a telephone, and an online survey experiment. *Field Methods*, *34*(3), 206–222.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*, *27*(1), 77–83.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

United States Census Bureau (2023). Population and housing unit estimates tables. https://www.census.gov/programs-surveys/popest/data/tables.2021.List_321237334.html#list-tab-List_321237334

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: a literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Vannieuwenhuyze, J. T., & Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: three methods to disentangle selection and measurement effects. *Sociological Methods & Research*, *42*(1), 82–104.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*, 409–422.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68.