A New Approach to Creating Population-Representative Data for Demographic Research

Abstract

The evaluation of innovative web-based data collection methods that are convenient for the general public and yield high-quality scientific information for demographic researchers has become critical. Web-based data collection methodologies are crucial for researchers with nationally-representative research objectives but without the resources of larger organizations. The web mode is appealing because it is inexpensive relative to in-person and telephone modes, and it affords a high level of privacy. We present estimates of population parameters related to reproductive health and family formation from a sequential mixed-mode web/mail data collection, conducted with a national probability sample of U.S. adults from 2020-2022. We compare these estimates to those obtained from a benchmark face-to-face national survey of population reproductive health: the 2017-2019 National Survey of Family Growth (NSFG). This comparison allowed for maximum design complexity, including a complex household screening operation (to identify households with persons aged 18-49). We demonstrate the ability of this national web-based data collection approach using address-based sampling and sequential mixed-mode design to 1) recruit a representative sample of U.S. persons aged 18-49; 2) replicate key survey estimates based on the NSFG; 3) reduce complex sample design effects relative to the NSFG; and 4) reduce the costs per completed survey.

Key Words: National Web Survey, Mixed-Mode Data Collection, Probability Sampling, Population Health, Reproductive Health

Introduction

The vast majority of all persons between the ages of 18 and 49 who live in the U.S. now use the internet (Vogels, 2021). The percentage of this population with internet access, via PCs or mobile devices, is approaching 95% (Pew Research Center, 2021), even across historically marginalized groups such as the poor (Couper et al., 2018). This increase in internet use has motivated rapid advances in applications of survey methodology to web-based data collection, revealing the many advantages of the web mode for population science (Biemer et al., 2022; Tourangeau et al., 2013). Perhaps the most important of these is the privacy with which respondents can complete a web survey. In either face-to-face or phone surveys, where interviewers are administering the survey content, interviewers may cause respondents to provide more socially desirable responses to sensitive questions (Sakshaug et al., 2010; Chang & Krosnick, 2009; Kreuter et al., 2008; Fricker et al., 2005). The privacy afforded by web surveys is crucial for demographic research on many potentially sensitive topics such as sex and reproductive health, romantic relationships, earnings, assets and remittances, or attitudes and beliefs. The web mode also facilitates the use of complex questionnaires relative to any form of paper questionnaire, including mailed surveys. In addition, respondents can access web surveys from virtually anywhere, which is crucial for any topics linked to respondent mobility, and web surveys have a much lower cost per interview than any other mode.

Despite these attractive features of the web mode for survey data collection, there are still important drawbacks to such an approach. First, no frame of internet users exists from which to draw a sample, so probability samples are generally selected from commercially-available lists

of addresses, and a different mode (e.g., mail) is used to recruit sampled households to participate in the study. Second, researchers have less control over within-household selection procedures if the objective of a study is to randomly select one eligible respondent from within a household. This generally requires a two-step process of initially requesting a sampled household to complete a screening questionnaire, and then randomly selecting an eligible person from all identified eligible persons in the household. Finally, web surveys are known to have lower response rates compared to non-web modes (Braekman et al., 2022; Daikeler et al., 2020; Tourangeau et al., 2013; Manfreda et al., 2008; Shih & Fan, 2008). Given these drawbacks, two crucial questions remain for demographic researchers:

(1) Can web surveys be used to produce statistically-efficient population estimates based on national probability samples in a cost-efficient manner?

(2) Are web surveys characterized by more selection bias than other modes?

In this study, we apply the advances in knowledge of the strengths and weaknesses of using the internet for scientific data collection to address these fundamental issues in the use of web surveys for demographic research. The survey methodology of maximizing response to data collection efforts under fixed constraints, while controlling the risk of nonresponse bias across measures within the same survey, has also evolved at a rapid pace, providing the tools required for conducting web-based data collection in a way that minimizes selection bias. The problem of reduced response rates in web surveys can be remedied in practice with sequential mixed-mode designs, where alternative non-web modes such as mail and telephone are used to follow up with non-respondents, increasing response rates while decreasing the nonresponse bias in estimates (Olson et al., 2021; Axinn et al., 2015; Tourangeau et al., 2014; Millar & Dillman, 2011). The

general approach of inviting national samples of households to participate in "push-to-web" data collections, via mailed invitations that ask one or more individuals from a sampled household to respond to a survey on the web, is still in its relative infancy. Some national surveys based on probability samples have started to explore the feasibility of using these types of sequential mixed-mode approaches that allow for the self-administration of survey content (Brick et al., 2011, 2012; Han et al., 2013; Montaquila et al., 2013; Zimmer et al., 2015; Biemer et al., 2016, 2018; DeBell, Amsbary, et al., 2018; DeBell, Maisel, et al., 2018; Olson et al., 2021; Luijkx et al., 2021). In general, these initial studies have found support for the use of such approaches and subsequently implemented them to varying extents.

Web surveys are a key feature of the future of data collection for many good reasons. This paper describes a significant effort to perfect the quality of the web survey approach for nationally representative demographic research. We present results from a web-based approach to conducting national survey data collection and compare them to results from a benchmark face-to-face national survey of population health: The National Survey of Family Growth (NSFG). The NSFG allows us to test maximum design complexity, including a complex household screening operation (necessary to identify the subset of households that contains eligible persons aged 18-49). We evaluate the ability of this national web-based data collection approach using address-based sampling and sequential mixed-mode design to 1) recruit a representative sample of U.S. persons between the ages of 18 and 49; 2) replicate key survey estimates based on the NSFG; 3) reduce complex sample design effects relative to the NSFG; and 4) reduce the costs per completed survey.

Background

The Increasing Inefficiency of In-Person Data Collection

A fundamental principle in the collection of data for research in the social sciences is that survey costs and errors are strongly and inversely connected (Groves, 2004). Reduction of error usually involves increasing costs, and reduction of costs usually involves increasing error. As the field of survey methodology has evolved, dozens of breakthroughs have identified specific tools that can be used to minimize changes in error while reducing costs. Breakthroughs related to the use of web surveys for samples of the general population are particularly important because web surveys reduce costs substantially relative to face-to-face surveys (Tourangeau et al., 2013). For example, research on web surveys demonstrates that careful use of a second mode of contact, such as telephone or mail, for a subset of respondents who are unable to participate (e.g., due to lack of access to the internet) or who choose not to participate via the web can significantly lower the nonresponse error in web surveys (Axinn et al., 2015; Bandilla et al., 2014; Vannieuwenhuyze, 2014). These design advances have now cumulated to make large-scale sequential mixed-mode designs a robust alternative to large-scale face-to-face surveys (Braekman et al., 2022; Couper, 2013).

The potential of web-based, sequential mixed-mode designs to yield statistical results consistent with more expensive face-to-face surveys is especially important given the large financial resources dedicated by federal funding agencies to survey data collection each year (Klausch, Hox, et al., 2015; Klausch, Schouten, et al., 2015; Presser & McCulloch, 2011). Survey data collection represents a significant expenditure for governments worldwide. For example, ten federal statistical agencies spent more than 1.3 billion dollars on surveys in the U.S. in 2004

(Presser and McCulloch, 2011; see also https://www.whitehouse.gov/omb/information-regulatory-affairs/statistical-programs-standards/), and federal budgets for statistical agencies have essentially remained flat since this time (Pierson, 2020). Although the U.S. government dedicates significant financial resources to fund these surveys each year, it's important to note that many other studies, such as randomized controlled trial (RCT) studies, also collect data using survey methods (e.g., Ginde et al., 2013; Formica et al., 2004; Meyers et al., 2003; Wilson et al., 2009). In addition, many other survey data collections are funded by the National Science Foundation (NSF), the National Institutes of Health (NIH), or private organizations, meaning that even more financial resources are dedicated to collecting survey data every year. The search for robust scientific methods of data collection that cost less than in-person surveys is also important outside the U.S. Research on data collection alternatives that produce high-quality scientific information in a more cost-efficient manner is clearly needed, especially in light of the COVID-19 pandemic that completely halted – at least temporarily -- in-person interviewing activities worldwide.

The rising costs of face-to-face survey data collection are of particularly high significance to all U.S. federally-supported research on the general population. In an era of growing survey nonresponse, the average effort required to complete a face-to-face survey interview – or the average "hours per interview (HPI)" an interviewer must work to complete each interview – has grown dramatically in the last 25 years. Survey management data from six years of the NSFG (Wagner et al., 2021) illustrate this trend in **Figure 1**. Even within the same survey, following the same design over time, the increase in HPI has been approximately 3.3% per year.

(Figure 1, about here)

Some of the largest components of this increase in HPI are lifestyle changes (multiple partners working, long commutes, etc.) that make it harder to find people at home, along with the general public's resistance to participating in survey interviews, which forces interviewers to spend more time finding respondents (often by driving significant distances), explaining the reasons for the interview request, and answering all questions raised by each respondent. Thus, the total costs of face-to-face survey interviewing for representative samples of the general population have grown several-fold. This means that the vast majority of demographic researchers working on smaller research teams are not able to conduct nationally representative population studies using interviewer-administered data collection.

These cost increases are not limited to face-to-face survey data collection. Between 1990 and 2015, the average hourly pay rate for a trained telephone survey interviewer increased from $7.77 to $19.34 (a 249% increase) on a comparable interviewing task (in the Panel Study of Income Dynamics, or PSID, per personal communication with PSID leadership), and this rate continues to increase. It would be impossible to continue absorbing this level of inflation within the fixed budget caps associated with many federally-funded projects. More cost-efficient alternatives that still yield high-quality data are needed.

The Evolution of Web Survey Tools

As internet access has spread throughout the population, conducting survey interviews via the web has become an attractive method for lowering data collection costs (Tourangeau et al., 2014;

Couper, 2008). Aside from the cost savings, the advantages of web surveys include portability, flexibility, and privacy – web surveys allow respondents to complete surveys at whatever time and location is convenient and private for them. These properties extend to multiple devices including personal computers, laptops, tablets, and smartphones, further providing respondents with more options for convenience with little difference in measurement error between the devices (given appropriate instrument design; e.g., Couper et al., 2016; Lugtig & Toepoel, 2015; Mavletova & Couper, 2013).

Unfortunately, the advantages of the web mode can be offset by the serious disadvantage of low response rates and the potential for both coverage bias and nonresponse bias to mislead investigators. Sampling frames of up-to-date email addresses are generally not available to researchers outside of specific contexts, and even if such a frame were available for a specific population, individuals without internet access would not be included in the frame, introducing a risk of coverage bias (Couper et al., 2018). This means that large-scale web surveys are often conducted via address-based sampling (to ensure high population coverage of the sampling frame) combined with mailed invitation letters inviting individuals to complete surveys online.

When carefully designed, web surveys have many positive characteristics. Some studies find that web surveys have less measurement error than surveys administered using other modes. This includes higher rates of reporting potentially sensitive information relative to telephone interviewing (Kreuter et al., 2008) and more accurate reporting on sensitive items involving undesirable characteristics (Chang & Krosnick, 2009; Fricker et al., 2005; Sakshaug et al., 2010). In general, the self-administered nature of web surveys consistently produces less response bias

due to social desirability than interviewer-administered modes (Kennedy et al., 2016; Lindhjem & Navrud, 2011; Holbrook & Krosnick, 2010). On the other hand, some studies indicate that web surveys may produce higher rates of "don't know" responses than interviewer-administered surveys (Heerwegh & Loosveldt, 2008). Measurement differences between face-to-face and web surveys need careful consideration at initial design stages depending on the survey content (Nielsen, 2011; Heerwegh, 2009). Careful design of the web content in a manner that maximizes data quality is therefore very important (Couper, 2008).

The focus of the present study is on one-time cross-sectional national surveys, the feasibility of replacing a large and complex face-to-face survey that includes a household screening operation (to determine eligibility) entirely with a web/mail survey, and using these data to estimate the health and fertility characteristics of the U.S. population aged 18-49.

Advances in Sequential Mixed-Mode Web Surveys

Recent research demonstrates that a sequential mixed-mode approach can compensate for the lower response rates known to be generated by web surveys and strengthen the web survey approach for robust, representative measurement (Axinn et al., 2015; Tourangeau et al., 2014). By focusing data collection effort with a second alternative mode (e.g., mail) on those cases that do not initially respond to web surveys, data collection can efficiently compensate for the web-specific noncoverage and nonresponse and create more robust statistics that more closely represent the full population (Millar & Dillman, 2011). Careful analysis of changes in key statistics as effort in a second mode is added can be used to adjust the amount of effort placed in a second mode and efficiently produce estimates with reduced bias.

Some studies suggest that when individuals are simply given a choice of responding by web or mail, the overall response rate is lower than when only a single response option is offered at a time (Medway & Fulton, 2012). Current research demonstrates that giving respondents a choice between web and mail may increase nonresponse in the absence of an extra incentive for completing the survey online (Biemer et al., 2016). For this reason, our proposed methodology features a *sequential* mixed-mode design, in which respondents are first asked to complete the survey over the web, and only when that invitation is not successful are they invited to complete the survey by mail. A prior analysis performed by the American National Election Studies (ANES) found that additional effort and follow-up applied to the mail/web sample were important for improving representativeness (DeBell, Amsbary, et al., 2018; DeBell, Maisel, et al., 2018), and we also found evidence of this in prior methodological evaluations (Wagner et al., 2022; West et al., Under Review).

Objectives of the Current Study

Our study focuses on the efficient conversion of a complex, lengthy, cross-sectional, national face-to-face survey of family formation and reproductive health to a web format for the collection of the same survey data from a national probability sample of persons between the ages of 18 and 49. No prior studies have attempted to convert a major national health survey like the NSFG into a web/mail format, and then employ a sequential mixed-mode design for both the screening and main data collection stages to field the fully self-administered survey in a cross-sectional national sample of households. The only analogous examples with national scope (in the U.S.) of which we are aware can be found in educational (Montaquila et al., 2013; Han et al.,

2013; Brick et al., 2012, 2011) or political (DeBell, Maisel, et al., 2018; DeBell, Amsbary, et al., 2018) contexts. Many investigators in the health sciences would benefit from this more cost-efficient approach to the large-scale collection of health survey data, but the approach has never been rigorously evaluated in this context.

The overall objectives of this study were four-fold:

1. Evaluate the features of the respondents recruited using this type of approach and the population that they would represent prior to any types of adjustments to base sampling weights;
2. Compare weighted survey estimates based on this web/mail approach to those computed from a benchmark national face-to-face survey (the NSFG) measuring the same content in a similar time frame;
3. Evaluate design effects on survey estimates due to the complex probability sampling designs employed in each case; and
4. Compare the data collection costs per completed survey associated with these two approaches.

Methodology

Overview of the American Family Health Study

We initiated data collection in April 2020 with a national address-based probability sample of more than 41,000 U.S. addresses. We called this new study the American Family Health Study (AFHS; see afhs.isr.umich.edu for additional details). The AFHS used a sequential mixed-mode mail/web protocol for push-to-web household (HH) screening to identify eligible persons aged

18-49. One eligible individual was then randomly selected from each HH with eligible persons present and invited (either by a mailed letter or email, if provided in the screening questionnaire) to complete a 60-minute web survey on the same reproductive health and family formation topics measured in the NSFG, using a second sequential mixed-mode mail/web protocol that encouraged the selected persons to respond to the "main" survey via the web. As part of this protocol, individuals who did not respond to the full 60-minute web survey were subsequently invited to complete a shorter paper questionnaire that was sent by mail. This reduced-length questionnaire primarily included items that were asked of all persons and did not require filter questions or complex skip logic. Conversion of the NSFG content to self-administered web and mail formats was not a trivial exercise; see Appendix II for more details on this process.

The data collection was split into two replicates, both of which were based on national samples. Data collection for the first sample continued until June 2021, and the second sample was fielded between September 2021 and April 2022. Additional details regarding the AFHS sample design and data collection methodology can be found elsewhere (see https://afhs.isr.umich.edu/about-the-study/afhs-methodology/).

AFHS Screening Protocol

The AFHS screening questionnaire was designed to collect a list of persons aged 18 years and over in the household. In the first phase of this protocol, we selected a stratified probability sample of addresses, oversampling addresses predicted to have an age-eligible (18-49 years old) person present (based on commercial data from Marketing Systems Group; see West et al., 2015) and located in high-density minority areas (based on data from the American Community

Survey). Sampled households received a mailed invitation (including a $2 cash incentive) addressed to the resident of a particular state, inviting an adult member of the household to complete a screening questionnaire online (available in English or Spanish). In the second phase of this protocol, a follow-up reminder was sent one week after the mailed invitation in the form of a postcard.

In the third phase of screening, a follow-up mailing that included a paper version of the screening questionnaire was sent two weeks after the postcard. In the fourth phase of screening, 28 days after the initial invitation, a random subsample of 5,000 non-finalized sampled addresses was sent a priority mailing with a final invitation to complete the screening questionnaire and an additional $5 incentive; this served to significantly increase the response rate to the screening invitation (Wagner et al., 2022). Information obtained from completed screening questionnaires was used to identify eligible persons within the sampled households. If there was only one eligible person in the household, or the person completing the screener was randomly selected to take part in the main survey, then that person was immediately invited to complete the main AFHS survey online. If there was more than one eligible person, one person was randomly selected and then invited -- either by email or via a mailed letter -- to complete the main survey.

AFHS Main Data Collection Protocol

Once an eligible respondent was randomly selected from a sampled household completing the screening questionnaire, the main data collection protocol was initiated. The main protocol differed slightly depending on whether the screener respondents were also selected to be the main respondents and whether they provided their emails or text-enabled phone numbers to be

contacted in the main stage. **Figure 2** summarizes the two alternative sequences of contact attempts and how they are divided into four distinct phases:

- **Phase 1:** An initial invitation to complete the main survey online was sent by email or mail to the selected respondent, and the letter promised a $70 token of appreciation once the completed survey was received. During this phase, respondents responded to the initial push-to-web invites without receiving any follow-up contact attempts. The main survey could be completed in English or Spanish.
- **Phase 2:** Two weeks later, selected cases who had not yet responded were followed up by either a postcard or email reminder (if the selected respondent provided an email address). For those for whom we had an email or text-enabled phone number, they received an additional email or text reminder in the third week. During this phase, follow-up contact attempts were made, but the costs of these attempts were relatively low.
- **Phase 3:** In the fourth week of follow-up, we mailed a substantially shortened paper version of the questionnaire but still encouraged the respondent to complete the survey online. For eligible nonrespondents for whom we did not have an email or text-enabled phone number, we mailed an additional paper questionnaire in the sixth week in a USPS priority mailer. For those for whom we had an email or phone number, we sent an email or text reminder in the fifth week. Therefore, during this phase, active nonrespondents were provided the additional option of responding by mailing back a completed version of the reduced questionnaire, even though they were still encouraged to respond via the web.
- **Phase 4:** After four or six weeks, our calling center staff made reminder telephone calls to nonrespondents with telephone numbers available from either commercial data sources

linked to our sampling frame or the initial screening questionnaire (83% of these nonrespondents had telephone numbers available – although some (12%) of these were found to be invalid during the reminder calls). These staff did not administer the survey, but rather encouraged the nonrespondents to self-administer the survey and provided any information that would assist them in doing so. During this phase, telephone reminders were the final attempts to convert nonrespondents, and these calling efforts served to significantly increase response rates (West et al., Under Review).

(Figure 2, about here)

The AFHS also embedded an experiment examining the effects of modular survey design (West et al., Under Review) during the main data collection stage. Household residents selected from the completed screening questionnaires were randomly assigned to either a full survey condition or a modular design condition. In the former condition, sampled individuals were asked to complete the entire 60-minute questionnaire in one sitting, and could take breaks and return to the survey if desired. In the latter condition, the questionnaire was divided into three modules of roughly equal length, and sampled individuals were invited to complete the three modules at their leisure, with a two-week break in-between the invitations to complete the modules. The data collection protocol used for each of the three modules is the same as depicted in **Figure 2**.

For this study, the full vs. modular design is considered a feature of the AFHS data, but not an analytical focus. Overall, we found that the modular design was not effective at increasing response and completion rates (West et al., Under Review) and that both the survey responses and socio-demographic measures collected from the modular design condition were quite similar

to those collected in the full-survey condition (see Appendix I). Thus, we combined the data from the two conditions and computed three sets of survey weights that combined the full-survey respondents with respondents to either module 1, modules 1 and 2, or modules 1, 2, and 3. These weights accounted for differential probabilities of selection, differential probabilities of completing both the screener and main surveys (either particular modules of the main survey, or the full survey), and calibration adjustments to known population totals from the American Community Survey. We then used these weights (where the weight used in the analysis depended on the module where a particular question was asked) to produce one set of AFHS estimates. The attrition due to the modular design resulted in variations in the sample sizes across AFHS items. Since respondents had to complete prior modules to proceed to the later modules, the sample sizes for items in module 1 were the largest, followed by module 2 and then module 3 (West et al., Under Review).

Response Rates

The first national sample replicate of the AFHS obtained an overall response rate in the screening stage of 15.0% and a conditional AAPOR RR4 response rate of 66.0% in the main stage. The second national sample replicate obtained a screening-stage response rate of 17.8% and a conditional main-stage response rate of 62.4%. For individuals randomly assigned to the modular condition in replicate 1, completing at least two sections of the questionnaire in the first 20-minute module was counted as a partial response. These two rates resulted in net AAPOR RR4 response rates of 9.9% and 11.1% for replicates 1 and 2, respectively. See Appendix III of the supplemental materials for detailed descriptions of these response rate calculations.

Analytic Approach

With our analyses, we sought to compare the 2020-2022 AFHS with the 2017-2019 NSFG (restricted to persons between the ages of 18 and 49) in terms of 1) respondent sample composition, 2) key survey estimates, 3) complex sample design effects, and 4) costs per completed case.

Respondent Sample Composition (Study Objective 1)

First, in terms of respondent sample composition, male and female respondents were compared separately in terms of their distributions on race and ethnicity, education, age, and marital status, using design-adjusted Rao-Scott chi-square tests. For AFHS, these demographic distributions were weighted by the *base sampling weights*, whereas for NSFG, the distributions were weighted by the *final sampling weights* (i.e., including nonresponse and calibration adjustments based on control totals provided by the U.S. Census Bureau). We did not consider nonresponse or calibration adjustments for the base AFHS sampling weights in this initial analysis so that we could evaluate the features of the sample respondents and the target population that they would represent. The fully-weighted NSFG distributions (which again were based on U.S. Census data) were therefore regarded as the benchmarks. We examined whether AFHS demographic distributions were approaching the NSFG benchmarks as the main data collection proceeded across the four phases in **Figure 2**, and we only considered the base sampling weights when computing AFHS estimates to examine sample composition.

Key Survey Estimates (Study Objective 2)

Second, in terms of key survey estimates, we identified 42 and 89 key measures in the male and female surveys, respectively (see the Figures in the Results section for details). These 131 measures, each capturing important data on critical domains of family reproductive and health behaviors, were selected based on the following criteria:

- The measures had acceptable variability in the response values, together with low rates of missing data;
- The measures were also analyzed in recent descriptive reports for males and females prepared by NCHS using NSFG data (see https://www.cdc.gov/nchs/nsfg/nsfg_products.htm);
- The measures were also collected in the condensed mail questionnaire used for nonresponse follow-up; and
- The measures have also been analyzed in previous studies using AFHS data (Axinn et al., 2021).

Both AFHS and NSFG estimates were weighted by the *final survey weights* (where again, the AFHS weight depended on the module in which a variable was located), and design-adjusted standard errors were computed for the weighted estimates accounting for the complex sampling features inherent to each study. In the NSFG, these sampling features included stratification, cluster sampling, and weighting, and Taylor Series Linearization was used for variance estimation (per NCHS guidelines[1]). In the AFHS, bootstrap replicate weights were used to 1) fully account for uncertainty in the adjustment of the base sampling weights for nonresponse, and 2) capture gains in efficiency of the estimates due to calibration of the weights to known

[1] https://www.cdc.gov/nchs/nsfg/nsfg_2017_2019_puf.htm#variance

population features (Valliant, 2004). These replicate weights fully captured the stratified sampling and weighting inherent to the AFHS design (Heeringa et al., 2017). The estimates based on the AFHS data were then compared with those based on NSFG using multiple approaches.

First, we computed 95% confidence intervals for the population parameters based on the AFHS estimates and determined the fraction of these intervals that covered the NSFG point estimates (i.e., the benchmarks). Second, recognizing the uncertainty in NSFG benchmark estimates, we used independent-sample t-tests to compare the AFHS and NSFG estimates. We determined what fraction of estimates had standardized differences that were more than two pooled standard errors (i.e., pooled-SE = $\sqrt{SE_{AFHS}^2 + SE_{NSFG}^2}$) away from zero. Third, we computed fully-weighted AFHS estimates in each of the four phases of the main data collection protocol, and we evaluated the changes in AFHS estimates across these four phases. In these analyses, our main question was whether AFHS estimates became more consistent with the NSFG estimates as more respondents were recruited across the four phases and whether this improvement in estimates was introduced by the sequential mixed-mode design or simply via the weighting adjustments described above. Finally, we found that 89% of AFHS respondents participated via the web mode, while the remaining 11% returned the shortened paper questionnaire. We compared fully-weighted AFHS estimates based on the full respondent sample to those based on the web subsample, specifically evaluating whether providing the shorter paper questionnaire brought in respondents with different characteristics and changed the survey estimates.

Complex Sample Design Effects (Study Objective 3)

Third, we computed the distributions of estimated complex sample design effects associated with each of the weighted estimates compared between NSFG and AFHS. These so-called design effects are specific to each survey estimate, and capture the inflation in the variance of each of the weighted estimates relative to a simple random sample of the same size, due to the complex sampling features associated with each design and the use of the final survey weights in estimation. We hypothesized that the AFHS sample design, which only involved stratification and weighting adjustments, would reduce design effects relative to the NSFG, which relied on a clustered area probability sample to save on the costs of data collection.

Costs per Completed Survey (Study Objective 4)

Finally, we compared the data collection cost per completed survey in each of the two studies (conditional on the fixed infrastructure and development costs associated with each project), focusing on the final four quarters of data collection in the 2010-2020 NSFG so that the costs per completed interview would be as comparable as possible.

Results

Objective 1: Comparisons of Respondent Sample Composition

Table 1 reports the weighted estimates of socio-demographic distributions based on the full set of respondents (to the full survey and the initial module). We reiterate that we use the base sampling weights only (without adjustments for nonresponse or calibration) when computing the distributions based on AFHS respondents.

(Table 1, about here)

Compared to the benchmark distributions from the NSFG, the AFHS approach recruited significantly more respondents that were non-Hispanic White and higher educated (for both males and females). The youngest group of female respondents (18-24) was under-represented in AFHS, compared to NSFG. These results are certainly not unique to the AFHS and are quite common in web surveys of large populations (Boas et al., 2020; Wells et al., 2019; Sha et al., 2017; Simmons & Bobo, 2015; Dillman et al., 2014; Tourangeau et al., 2013; Baker et al., 2010). The results imply that nonresponse adjustments to the base AFHS sampling weights may be needed to correct for potential biases in AFHS estimates due to the race/ethnicity and education response differentials. Whether there would be bias in survey estimates based on the AFHS due to these differentials depends on the associations of race/ethnicity and education with the measures of substantive interest collected in the AFHS (Kennedy et al., 2016).[2]

The adjustment of sampling weights to account for 1) differential nonresponse across subgroups, and 2) calibration of the adjusted weights to population control totals tends to introduce more variability in the final survey weights, and thus has the potential to increase estimated design effects of survey estimates (i.e., inflate the variance of the weighted estimates). We later illustrate that the variance inflation introduced by these weighting adjustments *does not*

[2] Efforts to adjust for potential nonresponse bias via survey weighting, especially for web surveys that tend to have lower response rates, have the potential to inflate the variance of survey estimates. Adaptive survey design strategies have the potential to correct some of this bias during data collection or as a part of the data collection strategy (Zhang, 2022; Peytchev et al., 2022; Rosen et al., 2014), and when combined with post-survey weighting may further increase the efficiency of estimates (Särndal & Lundquist, 2019; Zhang & Wagner, 2022). Novel adaptive design approaches for web/mail data collections using national probability samples certainly need additional research consideration.

consistently result in larger design effects than seen in the NSFG (where much of the variance inflation arises due to the cluster sampling performed).

We also evaluated changes in the demographic compositions of the AFHS respondents across the four phases of main data collection. The compositions were largely stable, but there was some evidence that AFHS distributions better matched those of the NSFG as data collection proceeded. Specifically, the proportion of "high school or less" male respondents changed from 13% at phase 1 to 20% by the end of phase 4, getting closer to the NSFG benchmark at 34%. Similarly, the proportions of 18-24, 25-34, and 35-49 male respondents changed from 19%, 37%, and 44% in phase 1 to 22%, 33%, and 45% by the end of phase 4, closely matching the NSFG benchmarks at 22%, 34%, and 45%. These results suggest that the later phases of data collection contributed to reducing discrepancies between AFHS and NSFG distributions.

Objective 2: Comparisons of Key Survey Estimates

Across all 131 male and female estimates (which we remind readers were based on the final adjusted survey weights for each study), 53% of the AFHS estimates had confidence intervals that covered the NSFG point estimates, and for 39% of estimates, the standardized difference of the estimates between the two studies was more than two pooled standard errors away from zero. In general, we found that the weighting adjustments applied to the AFHS estimates were slightly more effective at shifting the estimates closer to the NSFG benchmarks than the sequential mixed-mode design, although weighting tended to increase the standard errors (SE) of the estimates. Across the four phases, the base-weighted AFHS estimates differed significantly from the NSFG benchmark estimates 34%, 40%, 42%, and 46% of the time, whereas the final

weighted AFHS estimates differed significantly from the NSFG benchmark 33%, 40%, 38%, and 39% of the time.

In general, we found that the main utility of the four-phase sequential mixed-mode design was to increase the respondent yield (and therefore reduce the standard errors of the AFHS estimates) and that the telephone reminders were particularly effective in this regard (consistent with West et al., Under Review). Specific phases of the protocol did not necessarily produce cumulative estimates that were significantly closer to the NSFG benchmarks after applying the final adjusted weights.

In terms of how the reduced-length paper questionnaire affected survey estimates, we found that the weighted estimates based on the full sample were less likely to be biased than those based on the web subsample: 59% of the estimates had a smaller bias based on the full sample. In addition, and as expected, estimates based on the full sample had smaller standard errors due to the larger sample size. Thus, adding the paper mode helped to improve the quality of the survey estimates.

We, therefore, focus our comparisons on the final weighted estimates based on the full AFHS respondent sample.[3] The key survey estimates based on the AFHS and NSFG data, including

[3] Future work also needs to consider comparisons of *subgroup estimates* between the NSFG and web/mail approaches like the AFHS. Our next step is to compare estimates for sociodemographic subgroups and evaluate whether some subgroups were affected more heavily by the changes in data collection mode. A related challenge is the absence of data from minors under the age of 18 in the AFHS; the NSFG did measure minors between the ages of 15 and 17, where working with parents or guardians to obtain consent was possible given the face-to-face mode. Techniques for obtaining consent from minors when using strictly web/mail approaches will need future consideration and evaluation, given the importance of this subgroup to demographic research. In addition to comparisons of subgroup estimates, comparisons of estimates describing *relationships* between variables will also

95% confidence intervals for the parameters being estimated, are compared using dumbbell plots in **Figures 3-6** below.

(Figures 3 through 6, about here)

The four figures above illustrate the general consistency of the AFHS estimates with the NSFG estimates. We find that significant differences between the NSFG and the AFHS generally arise for measures that were likely affected by the COVID-19 pandemic.[4] For example, among males, the mean number of months working for pay in the past 12 months was lower in the AFHS compared to the NSFG (**Figure 5**), which may have been a function of the so-called "Great Resignation" from jobs for pay during the pandemic. Reports of the use of birth control and emergency contraception among females (**Figure 4**) and ever cohabitating among males (**Figure 3**) also declined significantly in the AFHS, possibly due to changing social circumstances introduced by the pandemic (Axinn et al., 2021). Reports of "excellent" health and ever being tested for HIV also declined significantly for both males and females in the AFHS (**Figures 3 and 4**), possibly reflecting poorer health during the pandemic and fewer in-person medical visits for clinical testing.

be needed in future work (e.g., Axinn et al., 2021) to support the large-scale adoption of this type of web/mail approach.

[4] We generally recognize the distinct possibility that the COVID-19 pandemic did affect some of the descriptive estimates considered in this study, contaminating comparisons between the AFHS and the NSFG. As a robustness check, we compared AFHS estimates based on replicate 1 (2020-2021) to those based on replicate 1 and 2 (2020-2022) to see whether estimates based on data collected after the start of the pandemic were stable over the course of two years. We found only minor shifts in estimates and no significant differences. Future attempts to perform additional comparisons with survey data that were also collected from national samples using web/mail approaches during the pandemic would be helpful for understanding whether the differences in selected estimates reported here were arising due to the pandemic or simply due to the change in data collection mode.

Objective 3: Comparisons of Design Effects due to Complex Sampling

The final AFHS estimates after the weighting adjustments had substantially lower complex sampling design effects than the NSFG estimates. This was expected due to the lack of cluster sampling in the AFHS sample design. **Figure 7** below shows the distributions of the estimated design effects computed for all of the estimates generated from each of the two studies. This figure illustrates the significant advantages of address-based sampling that only involves stratification and weighting for unequal probabilities of selection and nonresponse adjustment/calibration when using the web/mail approach for national data collection. The general distribution of the estimated design effects is shifted rather dramatically toward higher values for the NSFG, and most of the AFHS design effects are tightly clustered around 1.5 - 2.5. From a perspective of statistical efficiency, the AFHS approach can reduce the expected inflation in the variance of estimates due to complex sampling.

(Figure 7, about here)

Objective 4: Comparisons of Costs per Completed Survey

Our cost comparison conditions on the presence of an established data collection infrastructure for each project (e.g., programmed CAPI instruments for the NSFG, a converted and programmed web instrument for AFHS, development of respondent materials, interviewer training, etc.), meaning that we do not account for these costs and only consider the costs associated with actual data collection activities. These costs include interviewer/call center staff

hiring and training, interviewer management and support, travel (NSFG only), telephone charges, postage, printing, and respondent payments.

When analyzing all of the data collection costs from the four most recent quarters of the NSFG (with 5,731 completed surveys) and the first replicate of the AFHS (which was closest in time to the last four quarters of the NSFG, with 998 completed surveys), we find that the NSFG cost was about $717 per completed survey, while the AFHS cost was about $417 per completed survey, corresponding to a cost savings of about $300 per completed survey. There are thus significant cost savings associated with the AFHS approach. In these quarters of the NSFG, data collection costs were dominated by field/interviewer management (61% of total data collection costs), interviewer travel (24.7% of total data collection costs), and respondent incentives (7% of total data collection costs). The AFHS costs were largely dominated by respondent mailings (50% of the total data collection costs) and respondent incentives (29% of the total data collection costs).

Discussion

Summary of Results

This study demonstrates that a mixed-mode data collection approach employing web and mail modes of data collection exclusively with a national address-based sample, including a screening phase to identify households with age-eligible individuals for the study, was generally able to replicate estimates related to reproductive health and family formation from the in-person NSFG at a significantly lower cost per completed case and with greater statistical efficiency. Revisiting our four main objectives, we summarize our findings as follows:

1. The web/mail approach does tend to recruit more non-Hispanic White and higher-educated individuals to complete the survey, but this was not unique to this web/mail study, and careful weighting approaches can compensate for this potential source of nonresponse bias;
2. Most of the estimates produced were statistically similar to those produced by the NSFG, with some of the remaining estimates likely having shifts introduced by the COVID-19 pandemic;
3. Design effects on the variances of estimates due to complex sampling are generally a fraction of those found in the NSFG, largely owing to the absence of area cluster sampling in the AFHS approach; and
4. The cost per completed survey was about $300 less in the web/mail approach compared to the NSFG when considering all data collection activities.

Collectively, these findings suggest that applying this type of data collection approach to a national probability sample can yield significant efficiency benefits while producing data that are largely similar to those created by a face-to-face interviewing approach.

Demographic Research Potential of this New Approach

The application of state-of-the-art responsive survey design tools to web surveys, along with the correct application of weighting adjustments for differential nonresponse, can now produce nationally representative population estimates that are quite similar to those generated by face-to-face surveys at a fraction of the cost, in less time, with more confidentiality. This creates many new opportunities to advance population science. First, the general replication of measurement from the NSFG opens many new possibilities for extending population research on family and

fertility topics. The web approach can be successful even when a pandemic or other disruptive shock prevents face-to-face interviewing. The high privacy and confidentiality enable additions for more potentially sensitive content. The speed and efficiency of the web survey approach allow scientists to conduct family and fertility surveys harmonized to selected key NSFG measures, but with a wider range of innovative measures that are not possible to collect in the NSFG. The lower costs and statistical efficiency of this approach provide the means for more frequent measurement, larger sample sizes for a similar cost, or more targeted studies, each of which has the potential to significantly enhance family and fertility research.

Second, all of the same advantages apply to other areas of demographic research. Even as alternative forms of data on actual behaviors become more widely available (e.g., Paul et al. 2016), organic data on employment, school enrollment, purchases, health care use, or changes in address cannot be used to reflect an individual's plans, preferences, attitudes, or expectations that are likely to drive subsequent long-term decisions about work, education, consumption, health or migration. Individual-level self-reports are a crucial component of the population-scale prediction of change and variation in the factors shaping longevity, health, migration, wealth, and well-being. This new approach can provide an accurate representation of the full population while taking advantage of the cost, speed, and relative privacy of web surveys. The success of this approach is an important opportunity for many different areas of population research.

Third, this approach can also be optimized for studies focusing on specific subgroups of the population. The combination of a first-step screening questionnaire done at a very large scale, followed by an in-depth study of a specific topic, affords many new opportunities for population

science. Here we demonstrate age-based screening with oversampling of specific race/ethnicity subgroups to mimic the NSFG face-to-face design. Clearly, that approach can be used to examine other age groups of the population or other race/ethnicity subgroups. But the approach can also be used to screen for other key subgroups of the population. Those could be sexual or gender identity subgroups, physical ability subgroups, such as the disabled, or personal experience subgroups, such as the divorced population. In addition, because this approach features an address-based sample selection procedure, it can be adapted to general population studies of smaller geographic areas, such as states, counties, or municipalities. Together, this flexibility in the application of the design to subgroups of the population opens many new opportunities for general population research.

Ongoing methodological research to continuously improve this approach and adapt it to evolving circumstances in society will be required. We have identified some clear next steps in this program of methodological research. The advances from that body of research will make the approach presented here *even more successful* at representing the general population. Here, we show that a nascent version of this approach can produce estimates that essentially replicate a rigorously designed and executed face-to-face survey of the U.S. population. Future improvements in the approach will serve to make it even more valuable for population-scale research on a broad range of demographic topics.

References

Axinn, W. G., Gatny, H. H., & Wagner, J. (2015). Maximizing data quality using mode switching in mixed-device survey design: Nonresponse bias and models of demographic behavior. *Methods, Data, Analyses*, *9*(2), 163–184. https://doi.org/10.12758/mda.2015.010

Axinn, W. G., West, B. T., Schroeder, H., & Banchoff, E. (2021, December 13). *Pandemic babies: The social organization of daily life, sudden disruptions to social activities, and national evidence of disruption of trends in U.S. fertility behavior*. Max Planck Institute for Demographic Research: Pandemic Babies? The Covid-19 Pandemic and Its Impact on Fertility and Family Dynamics, Virtual. https://www.demogr.mpg.de/en/news_events_6123/pandemic_babies_the_covid_19_pandemic_and_its_impact_on_fertility_and_family_dynamics_9210

Bandilla, W., Couper, M. P., & Kaczmirek, L. (2014). The effectiveness of mailed invitations for web surveys and the representativeness of mixed-mode versus Internet-only samples. *Survey Practice*, *7*(4).

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*, *185*(3), 798–821. https://doi.org/10.1111/rssa.12750

Biemer, P. P., Murphy, J., Zimmer, S., Berry, C., Deng, G., & Lewis, K. (2018). Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, *6*(2), 240–261. https://doi.org/10.1093/jssam/smx015

Biemer, P. P., Murphy, J., Zimmer, S., Berry, J., Lewis, K., & Shaofen, D. (2016). *A test of web/PAPI protocols and incentives for the residential energy consumption survey*. Annual Conference of the American Association for Public Opinion Research.

Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*, *8*(2), 232–250. https://doi.org/10.1017/psrm.2018.28

Braekman, E., Demarest, S., Charafeddine, R., Drieskens, S., Berete, F., Gisle, L., Heyden, J. V. der, & Hal, G. V. (2022). Unit response and costs in web versus face-to-face data collection: Comparison of two cross-sectional health surveys. *Journal of Medical Internet Research*, *24*(1), e26299. https://doi.org/10.2196/26299

Brick, J. M., Montaquila, J. M., Han, D., & Williams, D. (2012). Improving response rates for Spanish speakers in two-phase mail surveys. *Public Opinion Quarterly*, *76*(4), 721–732. https://doi.org/10.1093/poq/nfs050

Brick, J. M., Williams, D., & Montaquila, J. M. (2011). Address-based sampling for subpopulation surveys. *Public Opinion Quarterly*, *75*(3), 409–428.

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.

Couper, M. P. (2008). *Designing effective web surveys*. Cambridge University Press.

Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, *7*(3), 145–156.

Couper, M. P., Antoun, C., & Mavletova, A. (2016). Mobile web surveys: A total survey error perspectve. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. C. Tucker, & B. West (Eds.), *Total Survey Error in Practice* (pp. 133–154). Wiley.

Couper, M. P., Gremel, G., Axinn, W., Guyer, H., Wagner, J., & West, B. T. (2018). New options for national population surveys: The implications of internet and smartphone coverage. *Social Science Research*, *73*, 221–235. https://doi.org/10.1016/j.ssresearch.2018.03.008

Daikeler, J., Bošnjak, 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. https://doi.org/10.1093/jssam/smz008

DeBell, M., Amsbary, M., Meldener, V., Brock, S., & Maisel, N. (2018). *Methodology Report for the ANES 2016 Time Series Study*. Stanford University and the University of Michigan.

DeBell, M., Maisel, N., Brader, T., & Meldener, V. (2018, June). *Nonresponse bias in a nationwide dual-mode survey*. 2018 International Total Survey Error Workshop, Durham, NC. https://www.demogr.mpg.de/en/news_events_6123/pandemic_babies_the_covid_19_pandemic_and_its_impact_on_fertility_and_family_dynamics_9210

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Formica, M., Kabbara, K., Clark, R., & McAlindon, T. (2004). Can clinical trials requiring frequent participant contact be conducted over the Internet? Results from an online

randomized controlled trial evaluating a topical ointment for herpes labialis. *Journal of Medical Internet Research*, *6*(1), e6.

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

Ginde, A. A., Sullivan, A. F., Bernstein, S. L., Camargo Jr, C. A., & Boudreaux, E. D. (2013). Predictors of successful telephone contact after emergency department-based recruitment into a multicenter smoking cessation cohort study. *Western Journal of Emergency Medicine*, *14*(3), 287.

Groves, R. M. (2004). *Survey errors and survey costs*. John Wiley & Sons.

Han, D., Montaquila, J. M., & Brick, J. M. (2013). An evaluation of incentive experiments in a two-phase address-based sample mail survey. *Survey Research Methods*, *7*(3), 207–218.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied Survey Data Analysis* (Second). Chapman and Hall / CRC Press.

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

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846.

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37–67.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating Online Nonprobability Surveys* (No. 61). Pew Research Center.

Klausch, T., Hox, J., & Schouten, B. (2015). Selection error in single-and mixed mode surveys of the Dutch general population. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(4), 945–961.

Klausch, T., Schouten, B., & Hox, J. J. (2015). Evaluating bias of sequential mixed-mode designs against benchmark surveys. *Sociological Methods & Research*. http://journals.sagepub.com/doi/full/10.1177/0049124115585362

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

Lindhjem, H., & Navrud, S. (2011). Are Internet surveys an alternative to face-to-face interviews in contingent valuation? *Ecological Economics*, *70*(9), 1628–1637.

Lugtig, P., & Toepoel, V. (2015). The use of PCs, smartphones, and tablets in a probability-based panel survey effects on survey measurement error. *Social Science Computer Review*, *34*(1), 78–94.

Luijkx, R., Jónsdóttir, G. A., Gummer, T., Ernst Stähli, M., Frederiksen, M., Ketola, K., Reeskens, T., Brislinger, E., Christmann, P., Gunnarsson, S. Þ., Hjaltason, Á. B., Joye, D., Lomazzi, V., Maineri, A. M., Milbert, P., Ochsner, M., Pollien, A., Sapin, M., Solanes, I., … Wolf, C. (2021). The European Values Study 2017: On the Way to the Future Using Mixed-Modes. *European Sociological Review*, *37*(2), 330–346. https://doi.org/10.1093/esr/jcaa049

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., Vehovar, V., & Berzelak, N. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *Journal of the Market Research Society*, *50*(1), 79–104.

Mavletova, A., & Couper, M. P. (2013). Sensitive topics in PC web and mobile web surveys: Is there a difference. *Survey Research Methods*, *7*(3), 191–205.

Medway, R. L., & Fulton, J. (2012). When more gets you less: A meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, *76*(4), 733–746.

Meyers, K., Webb, A., Frantz, J., & Randall, M. (2003). What does it take to retain substance-abusing adolescents in research protocols? Delineation of effort required, strategies undertaken, costs incurred, and 6-month post-treatment differences by retention difficulty. *Drug and Alcohol Dependence*, *69*(1), 73–85.

Millar, M. M., & Dillman, D. A. (2011). Improving response to web and mixed-mode surveys. *Public Opinion Quarterly*, *75*(2), 249–269. https://doi.org/10.1093/poq/nfr003

Montaquila, J. M., Brick, J. M., Williams, D., Kim, K., & Han, D. (2013). A study of two-phase mail survey data collection methods. *Journal of Survey Statistics and Methodology*, *1*(1), 66–87.

Nielsen, J. S. (2011). Use of the Internet for willingness-to-pay surveys: A comparison of face-to-face and web-based interviews. *Resource and Energy Economics*, *33*(1), 119–129.

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., McCarthy, J. S., O'Brien, E., Opsomer, J. D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., & Wagner, J. (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. https://doi.org/10.1093/jssam/smz062

Pew Research Center. (2021). *Internet/Broadband Fact Sheet*. Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Peytchev, A., Pratt, D., & Duprey, M. (2022). Responsive and Adaptive Survey Design: Use of Bias Propensity During Data Collection to Reduce Nonresponse Bias. *Journal of Survey Statistics and Methodology*, *10*(1), 131–148. https://doi.org/10.1093/jssam/smaa013

Pierson, S. (2020). *FY20 Budget Brings Increases for NIH, Select Statistical Agencies*. Amstat News. https://magazine.amstat.org/blog/2020/04/01/fy20-budget-increases/

Presser, S., & McCulloch, S. (2011). The growth of survey research in the United States: Government-sponsored surveys, 1984–2004. *Social Science Research*, *40*(4), 1019–1024.

Rosen, J. A., Murphy, J., Peytchev, A., Holder, T., Dever, J. A., Herget, D. R., & Pratt, D. J. (2014). Prioritizing Low Propensity Sample Members in a Survey: Implications for Nonresponse Bias. *Survey Practice*, *7*(1), 1–8. https://doi.org/10.29115/SP-2014-0001

Sakshaug, J., Couper, M. P., & Ofstedal, M. B. (2010). Characteristics of physical measurement consent in a population-based survey of older adults. *Medical Care*, *48*(1), 64–71. https://doi.org/10.1097/mlr.0b013e3181adcbd3

Särndal, C.-E., & Lundquist, P. (2019). An assessment of accuracy improvement by adaptive survey design. *Survey Methodology*, *45*(2), 317–338.

Sha, M., McAvinchey, G., Quiroz, R., & Moncada, J. (2017). Successful Techniques to Recruit Hispanic and Latino Research Participants. *Survey Practice*, *10*(3), 1–9. https://doi.org/10.29115/SP-2017-0014

Shih, T.-H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, *20*(3), 249–271.

Simmons, A. D., & Bobo, L. D. (2015). Can Non-full-probability Internet Surveys Yield Useful Data? A Comparison with Full-probability Face-to-face Surveys in the Domain of Race

and Social Inequality Attitudes. *Sociological Methodology*, *45*(1), 357–387. https://doi.org/10.1177/0081175015570096

Tourangeau, R., Conrad, F., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

Tourangeau, R., Edwards, B., Johnson, T. P., Bates, N., & Wolter, K. M. (2014). *Hard-to-survey populations*. Cambridge University Press.

Valliant, R. (2004). The Effect of Multiple Weighting Steps on Variance Estimation. *Journal of Official Statistics*, *20*(1), 1–18.

Vannieuwenhuyze, J. (2014). On the relative advantage of mixed-mode versus single-mode surveys. *Survey Research Methods*, *8*(1), 31–42.

Vogels, E. A. (2021). *Digital divide persists even as Americans with lower incomes make gains in tech adoption*. Pew Research Center. https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/

Wagner, J., Guyer, H., & Evanchek, C. (2021). Using Time Series Models to Understand Survey Costs. *Journal of Survey Statistics and Methodology*, *9*(5), 943–960. https://doi.org/10.1093/jssam/smaa024

Wagner, J., West, B. T., Couper, M. P., Zhang, S., Gatward, R., Nishimura, R., & Saw, H.-W. (2022). An Experimental Evaluation of Two Approaches for Improving Response to Household Screening Efforts in National Mail/Web Surveys. *Journal of Survey Statistics and Methodology*, smac024. https://doi.org/10.1093/jssam/smac024

Wells, B. M., Hughes, T., Park, R., CHIS Redesign Working Group, & Ponce, N. (2019). *Evaluating the California Health Interview Survey of the Future: Results from a*

Statewide Pilot of an Address-Based Sampling Mail Push-to-Web Data Collection. UCLA Center for Health Policy Research.

West, B. T., Zhang, S., Wagner, J., Gatward, R., Saw, H., & Axinn, W. G. (Under Review). *Methods for improving the quality of national household surveys* [Paper submitted for publication, October 2022].

Wilson, C. L., Cohn, R. J., Johnson, K. A., & Ashton, L. J. (2009). Tracing survivors of childhood cancer in Australia. *Pediatric Blood & Cancer*, *52*(4), 510–515.

Zhang, S. (2022). Benefits of Adaptive Design Under Suboptimal Scenarios: A Simulation Study. *Journal of Survey Statistics and Methodology*, *10*(4), 1048–1078. https://doi.org/10.1093/jssam/smab051

Zhang, S., & Wagner, J. (2022). The Additional Effects of Adaptive Survey Design Beyond Post-Survey Adjustment: An Experimental Evaluation. *Sociological Methods & Research*, 004912412210995. https://doi.org/10.1177/00491241221099550

Zimmer, S., Biemer, P. P., Kott, P. S., & Berry, C. (2015). Testing a Model-Directed, Mixed Mode Protocol in the RECS Pilot Study. *2015 Proceedings of the Federal Committee on Statistical Methodology*.

Table 1. Estimated demographic distributions based on selection weights only in the AFHS (2020-2022) and fully-adjusted survey weights in the NSFG (2017-2019).[5]

	Female						Male					
	AFHS		NSFG		AFHS vs. NSFG		AFHS		NSFG		AFHS vs. NSFG	
	n	%	n	%	Pairwise p-value	Chi-squared Test of independence	n	%	n	%	Pairwise p-value	Chi-squared Test of independence
	1357		5557				989		4608			
Race/ethnicity												
Hispanic	223	16%	1505	20%	< 0.01		163	15%	1148	22%	< 0.01	
Non-H White	782	63%	2478	56%	< 0.01	$\chi^2(3) = 38.66$,	607	64%	2208	56%	< 0.01	$\chi^2(3) = 46.15$
Non-H Black	211	10%	1075	13%	< 0.01	$p < 0.01$	106	9%	792	12%	0.51	$p < 0.01$
Non-H Other	138	10%	499	10%	0.78		110	12%	460	11%	0.41	
Education												
High school or less	235	16%	1810	28%	< 0.01		195	20%	1793	34%	< 0.01	
Some college	407	30%	1927	34%	< 0.01	$\chi^2(3) = 197.87$	283	30%	1487	34%	0.07	$\chi^2(3) = 180.32$
4-year college	318	25%	875	19%	< 0.01	$p < 0.01$	236	24%	714	17%	< 0.01	$p < 0.01$
College+	394	30%	945	19%	< 0.01		275	27%	610	15%	< 0.01	
Age												
18-24	236	17%	1200	21%	0.02	$\chi^2(2) = 16.87$	189	22%	1103	22%	0.86	$\chi^2(2) = 0.10$
25-34	494	37%	2060	33%	0.02	$p = 0.02$	345	33%	1618	34%	0.90	$p = 0.98$
35-49	627	46%	2297	46%	0.73		455	45%	1887	45%	0.96	
Marital Status												
Married	532	46%	1912	42%	0.04		381	45%	1463	42%	0.26	
With partner	149	13%	751	15%	0.27		102	12%	521	13%	0.58	
Widowed	4	0%	34	1%	0.30	$\chi^2(3) = 10.63$	3	0%	14	0%	0.75	$\chi^2(3) = 5.18$
Divorced	103	6%	409	6%	0.46	$p = 0.19$	50	3%	264	4%	0.17	$p = 0.47$
Separated	36	2%	200	2%	0.72		19	1%	110	2%	0.59	
Never married	530	32%	2240	33%	0.50		432	38%	2230	39%	0.81	

[5] We remind readers that the AFHS estimates are ***not*** calculated using final weights for the purpose of this specific analysis. Correct construction of weights for differential non-response is a key element of producing population-representative estimates, as explained and demonstrated in subsequent analyses.

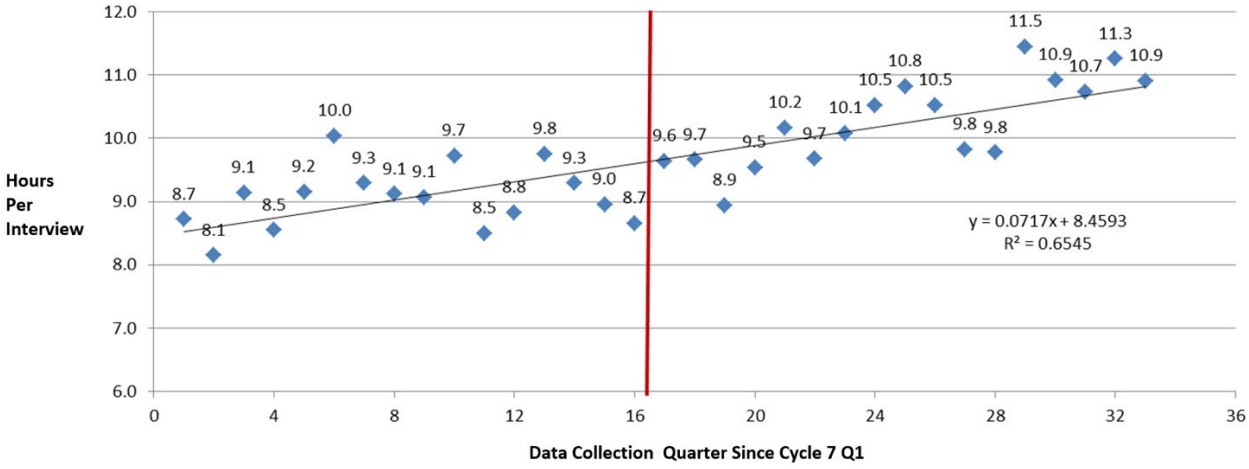


Figure 1. Hours per interview by quarter (from Cycle 7 Q1 to Cycle 8 Q17).

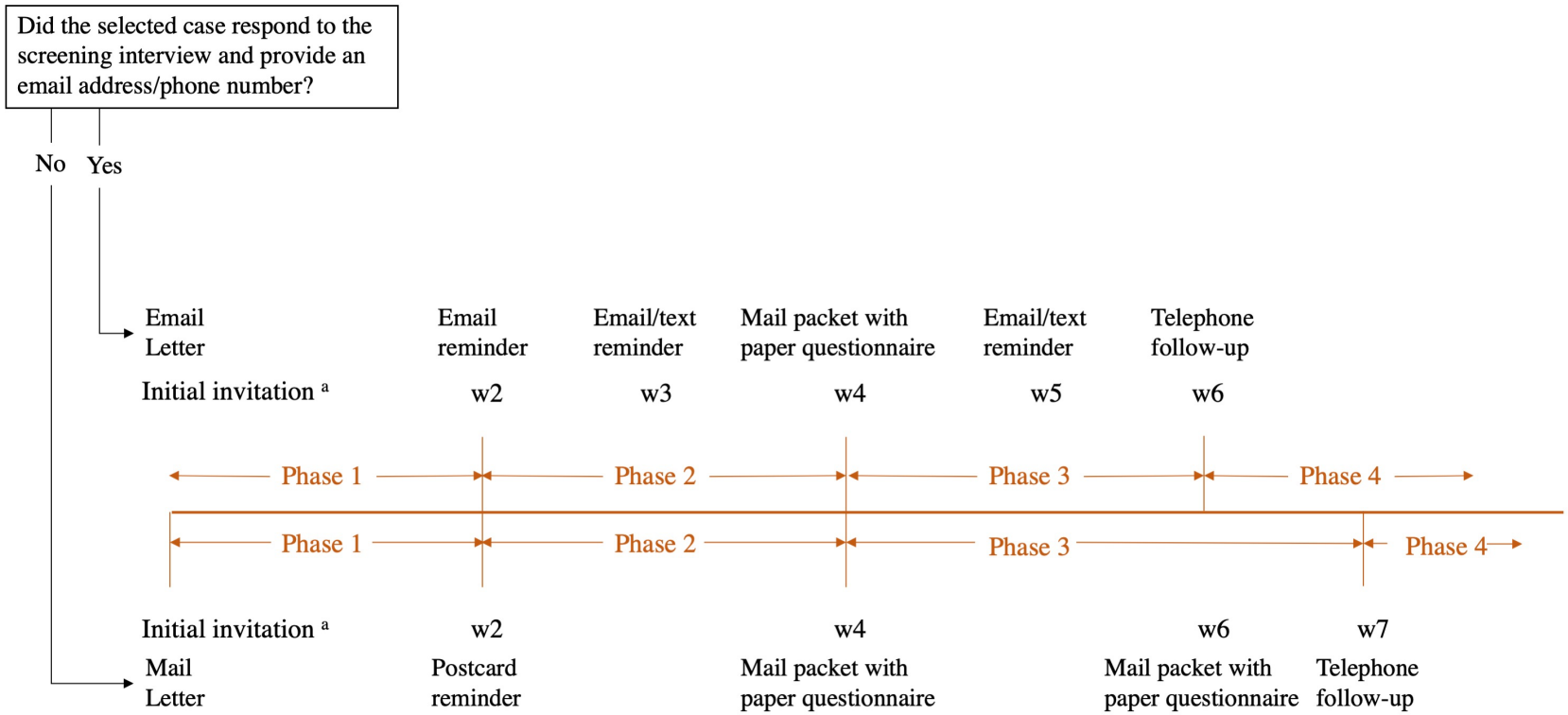


Figure 2. Contact attempts for sampled cases who were screener respondents (or not) and phases of the contact attempts.

Note: This figure shows the contact protocol for the Replicate 1 sample; the protocol for Replicate 2 was slightly different because we also collected contact information for the selected persons that were *not* the screening respondents and contacted them via email or text messages.

[a] The initial invitation either immediately followed the screening interview for the screener respondents who were selected for the main interview, or was scheduled 14 days after the screener interview for the other selected household members. Partial respondents were treated as nonrespondents in each phase.

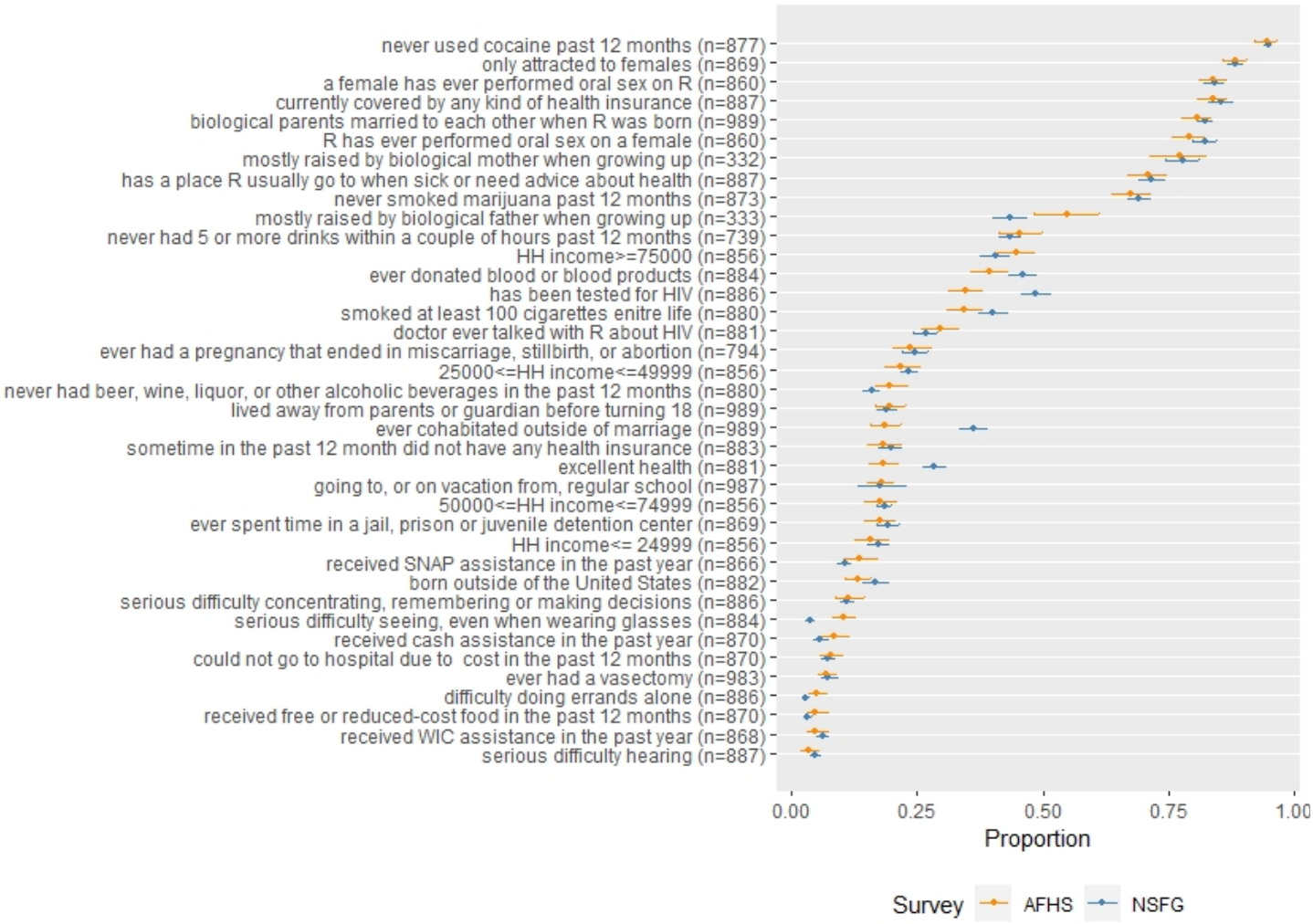


Figure 3. Estimated proportions based on AFHS (sample sizes in parentheses) and NSFG data on male respondents.[6]

[6] Estimates associated with substantially reduced sample sizes in Figures 3-6 are based on variables that were only measured on a subsample. For example, the item about mostly being raised by a biological father when growing up was only measured for persons who did not live with their biological or adoptive parents until they were 18 years old.

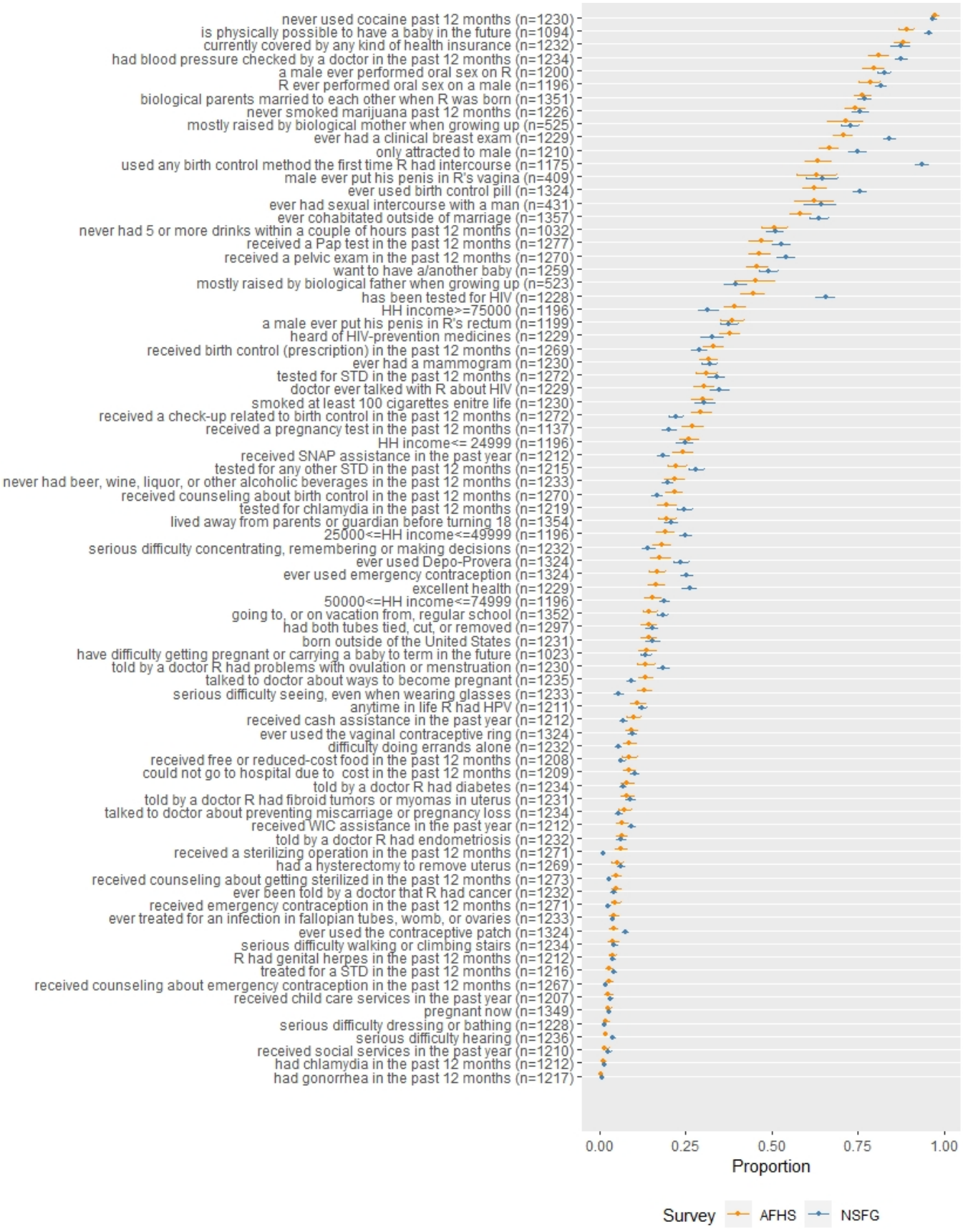


Figure 4. Estimated proportions based on AFHS (sample sizes in parentheses) and NSFG data on female respondents.

Figure 5. Estimated means based on AFHS (sample sizes in parentheses) and NSFG data on male respondents.

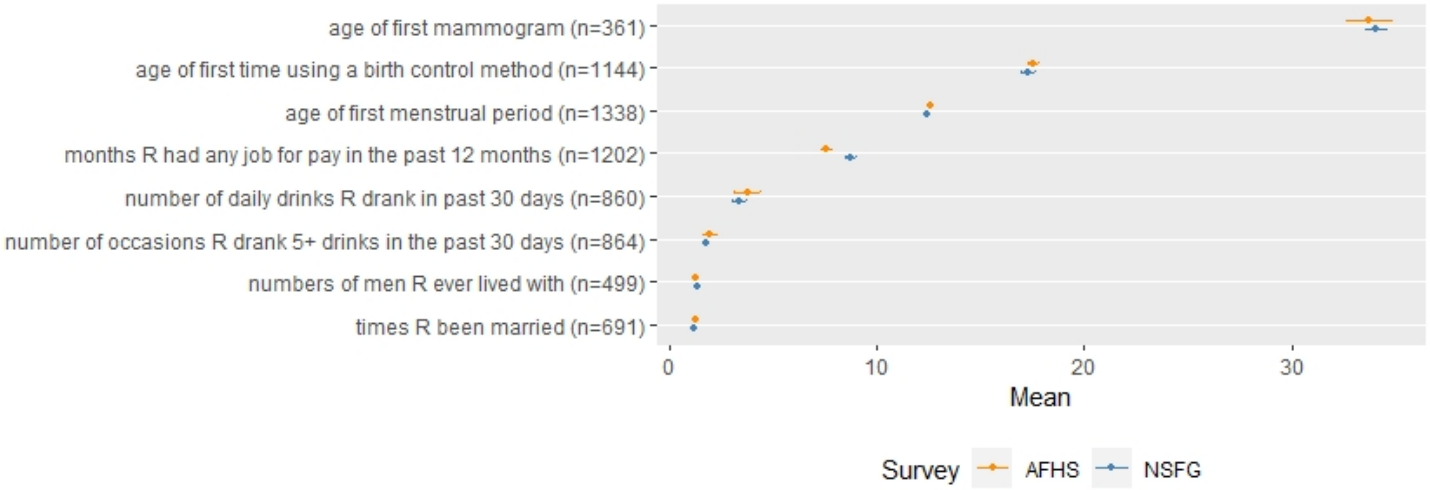


Figure 6. Estimated means based on AFHS (sample sizes in parentheses) and NSFG data on female respondents.

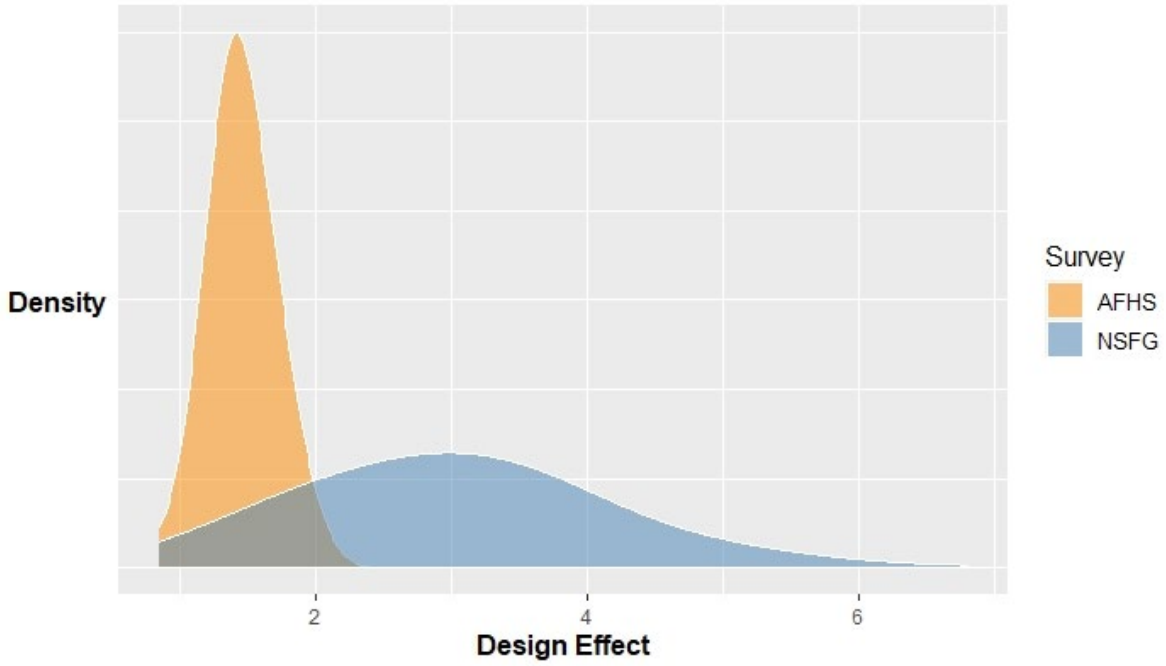


Figure 7. Distributions of estimated design effects for all estimates compared between the AFHS (orange/light shading) and the NSFG (blue/darker shading).

Supplementary Materials

Appendix I: Comparing Demographics of the Full- and Modular-Condition Respondents

Appendix II: Conversion of the NSFG Content to Web and Mail Formats

Appendix III: Response Rate Calculations

Appendix I: Comparing Demographics of the Full- and Modular-Condition Respondents

Table I.1 compares the demographic compositions of the full-condition and modular-condition respondents for females and males, respectively, based on replicate 1 data. As there are no statistically significant differences in the demographics, we combined the full and modular respondents into one sample for the analysis reported in the main text.

Table I.1. Demographic composition of female respondents of the full and modular condition.

	Female					Male				
	Full Condition		Modular		Full vs. Module	Full Condition		Modular		Full vs. Module
	n	%	n	%		n	%	n	%	
Sample size	297		278			211		207		
Race/ethnicity										
Hispanic	51	15%	49	17%	Chi2= .32,	35	15%	32	11%	Chi2= 7.47,
Non-H White	181	67%	166	65%	df= 3,	124	63%	135	65%	df= 3,
Non-H Black	39	10%	38	10%	p= .967	20	6%	17	12%	p= .195
Non-H Other	25	8%	25	8%		32	16%	23	11%	
Education										
High school or less	49	15%	46	17%	Chi2= 5.11,	41	19%	36	16%	Chi2 = 8.20,
Some college	82	27%	89	32%	df= 3,	61	28%	72	40%	df= 3,
4-year college	72	22%	59	25%	p= .317	57	27%	41	20%	p= .134
College+	94	35%	84	27%		52	26%	58	24%	
Age										
18-24	47	14%	60	21%	Chi2= 6.21,	57	32%	42	25%	Chi2= 2.12,
25-34	109	37%	102	38%	df= 2,	72	31%	71	33%	df= 2,
35-49	141	49%	116	41%	p= .106	82	37%	94	41%	p= .549
Marital Status										
Married	120	47%	118	51%	Chi2= 2.81,	79	42%	77	43%	Chi2= 0.49,
With partner	37	16%	27	11%	df= 3,	21	11%	21	11%	df= 3,
Widowed	1	0%	1	0%	p= .576	2	0%	0	0%	p= .944
Divorced	19	6%	24	5%		9	3%	10	2%	
Separated	6	2%	8	2%		4	1%	4	1%	
Never married	114	29%	99	30%		95	42%	95	43%	

Note: for Martial Status, the categories widow, divorced, and separated are combined for the Chi-squared test.

Appendix II: Conversion of the NSFG Content to Web and Mail Formats

A key step in achieving the aims of this study was converting the questionnaire content from the Computer Assisted Personal Interviewing (CAPI) format currently used by the NSFG to a web survey format. Our objective was the replication of all key content domains and the replication of nearly every specific question. That is, we sought to replicate almost all of the NSFG content.

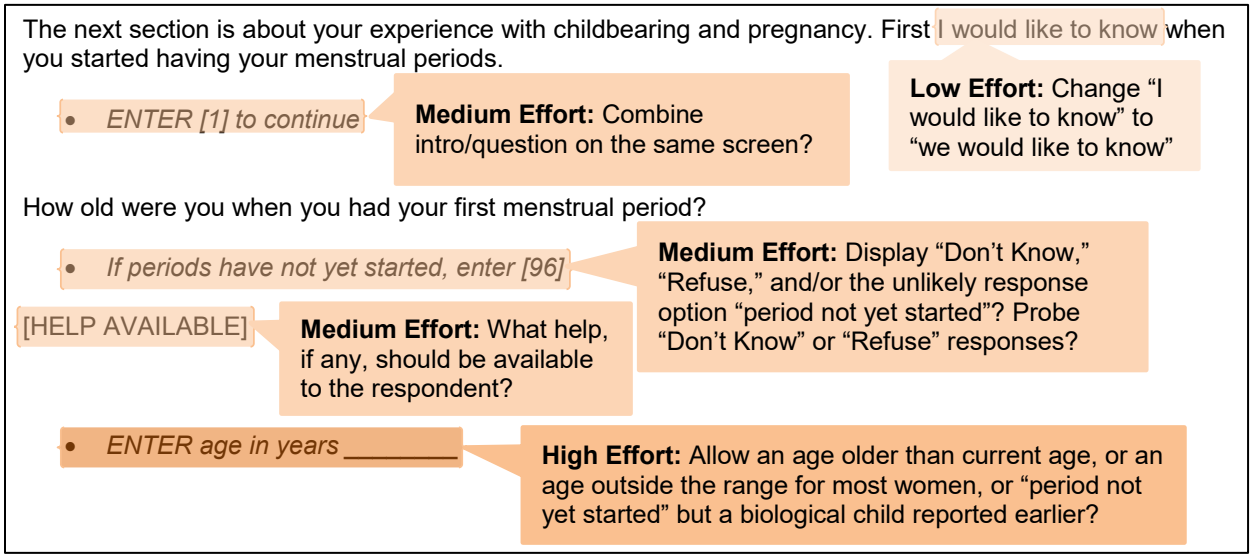
The task itself varied across content domains within the NSFG. **Table II.2** summarizes the content domains of the 2010-2020 NSFG female and male interviews. A portion of the NSFG interview is already conducted in Audio-Computer-Assisted Self-Interviewing (ACASI). Conversion of this content to a web survey format did not require as much effort as the conversion of the face-to-face content, as the design of the ACASI instrument already has much in common with the design of web-survey instruments. For other portions of the NSFG interview, the task became more complex, and the web survey also needed to be optimized for smartphone use. The target population of both the NSFG and this research (ages 18-49) will have a relatively high probability of attempting to complete the various modules on smartphones.

Table II.2. NSFG 2010-2020 Content Domains

Section	Female Questionnaire	Male Questionnaire
A	Household Roster; Childhood Background; Demographic Status	Household Roster; Childhood Background; Demographic Status
B	Pregnancy & Birth History; Adoption & Nonbiological Children	Sex Experience, Sexual Partners, Sterilization, Biological Children
C	Marriage and Cohabitation History; First Intercourse; Sexual Partners	Current Wife or Cohabitating Partner
D	Sterilization and Impaired Fecundity	Recent (Or Last) Sexual Partner(s) and First Sexual Partner
E	Contraceptive History and Pregnancy Wantedness	Former Wives and First Cohabitating Partner
F	Family Planning and Medical Services	Other Biological Children, Adopted Children, Pregnancies
G	Birth Desires and Intentions	Fathering
H	Infertility Services; Reproductive Health; HIV Testing	Desires and Intentions for Future Children
I	Insurance; Residence; Religion; Work; Child Care; Attitudes	Health Conditions and Health Services
J	Audio CASI: Abortion, Sexual Assault, Risky Behaviors	Residence; Religion; Military service; Work; Attitudes
K		Audio CASI: Abortion, Sexual Assault, Risky Behaviors

Some changes required low effort; for example, revising text with first-person pronouns. Other changes required moderate effort; for example, questions had to be reviewed to determine how to present less desirable response options without encouraging them (e.g., "Don't Know," "Refuse"), or acceptable response options not offered initially by the CAPI version (e.g., "If volunteered" response options). High -effort changes included determining how hard and soft edit checks should be converted into the web survey version and converting complex skips/routing to the paper (mail) backup. A hard edit check requires a respondent to correct a response that is impossible based on a response to a prior question or correct the response to the prior question. A soft edit check allows an unlikely response to be entered, but first checks with the respondent that their response was entered correctly. These are just a few examples of the more common decisions to be made and the effort involved in converting each survey question from CAPI to the web. Even a minor change to wording or the presentation of survey questions or response options has the potential to affect responses, so each NSFG question had to be carefully reviewed and changes were made on a case-by-case basis. **Figure II.1** below illustrates the decisions that were required to convert the NSFG's "age at menarche" survey question from CAPI to the web (as an example). Even "low effort" changes were labor-intensive when multiplied across many questions.

Figure II.1. CAPI to web translation issues for an example NSFG survey question.



As was noted above, another key challenge was replicating the NSFG contraceptive use calendar in the web survey format. NSFG respondents are asked to report sex and contraceptive use, on a month-by-month basis, for a retrospective period of up to 54 months. This procedure places a tremendous recall burden on respondents, is time-consuming, and likely creates substantial reporting error. The problem is exacerbated for those transitioning in and out of sexual relationships, and those who use temporary contraceptive methods such as oral contraceptive pills or condoms. The NSFG requires this long recall period to measure a large number of events and provide high statistical power for contraceptive failure analyses. The substantially cheaper web-survey approach that we describe could ultimately achieve the required power (person-months of exposure) by collecting more brief retrospective reports (which are less prone to misreporting) from a larger number of people. One thousand (1,000) NSFG interviews generate 54,000 person-months of exposure, but by using a 24-month reporting period among 5,000 people, we could generate 120,000 person-months of exposure with less likelihood of recall error for about the same cost. We also built in memory cues so that respondents could answer the

questions more accurately and quickly, and generated a visual summary of the event history calendar responses provided, allowing respondents to check their answers. See West et al. (2022) for more details on the development of the web-based contraceptive use calendar.

Although we considered other new tools for web survey data collection for this project, such as audio-narration of web interviews, extension to new interviewing interfaces increases the risk of the innovative approach that we implemented, detracting from our primary objectives. Given the state of the existing research, there is an important need to understand the effectiveness of translating a large and complex face-to-face survey into a standard, cost-efficient web survey instrument (which was programmed using state-of-the-art Blaise software), and evaluate whether a web-based approach can produce estimates with similar quality.

Reference

West, B.T., Axinn, W.G., Couper, M.P., Gatny, H., and Schroeder, H. (2022). A Web-Based Event History Calendar Approach for Measuring Contraceptive Use Behavior. *Field Methods*, 34(1), 3-19

Appendix III: Response Rate Calculations

This appendix explains how the response rates reported in the main text were calculated.

Screening Response Rate

The main text reports that the response rate of the screening phase was 15.0%. The screening response rate was calculated as

$$\frac{Elig + Inelig}{All - NS - (UH)(1 - e2)}$$

where:

Elig=Screened eligible

Inelig=Screened ineligible

All=All released sample

NS=Screened non-sample, including mail returned as vacant or no such number or no such street.

UH=Unknown household

e_2= 88.94%, the estimated occupancy rates of the UH cases. The estimation of e_2 is explained below.

The counts of AFHS replicate 1 sample units in the AAPOR response rate categories are summarized in the **Table III.1** below.

Table III.1. AFHS replicate 1 sample unit counts in the AAPOR response rate categories.

ScreenOutcome	AAPORCategory	Known	Frequency
ScreenNS	NS	Non-HH	668
ScreenOther_Final	UH		15137

ScreenOther_Final	UO	HH	1020
ScreenedEligible	I	HH	801
ScreenedEligible	O	HH	435
ScreenedEligible	P	HH	197
ScreenedEligible	R	HH	80
ScreenedIneligible	Inelig	HH	1043
Total			19381

Among the cases which household status was *known*, the occupancy rate was

$$e_1 = \frac{1020 + 801 + 435 + 197 + 80 + 1043}{1020 + 801 + 435 + 197 + 80 + 1043 + 668} = 84.26\%$$

This observed occupancy rate is different from the 87.91% estimated occupancy rate based on ACS 2019 one-year data. We assume that this difference comes from the 15137 cases in which household status was unknown (UH). By matching the AFHS full sample occupancy rate to the ACS occupancy

$$\frac{1020 + 801 + 435 + 197 + 80 + 1043 + 15137 * e_2}{19381} = 87.91\%,$$

we solved that the estimated occupancy rate of the UH cases e_2 is 88.94%.

Main Response Rates

The main text reports that the response rate of the main phase was 66.0%, conditional on participation in the screening phase. The main response rates were calculated as

$$\frac{I + P}{Elig}$$

where:

I=Interview

P=Partial

Elig=Screened Eligible

Overall Response Rates (RR4)

The overall response rates reported in the main text are obtained by the product of the screening response rate and main response rate.