*Journal of Survey Statistics and Methodology (2024) **12**, 932–960*

# PROXY SURVEY COST INDICATORS IN INTERVIEWER-ADMINISTERED SURVEYS: ARE THEY ACTUALLY CORRELATED WITH COSTS?

JAMES WAGNER *
LENA CENTENO
RICHARD DULANEY
BRAD EDWARDS
Z. TUBA SUZER-GURTEKIN
STEPHANIE COFFEY

Survey design decisions are—by their very nature—tradeoffs between costs and errors. However, measuring costs is often difficult. Furthermore, surveys are growing more complex. Many surveys require that cost information be available to make decisions during data collection. These complexities create new challenges for monitoring and understanding survey costs. Often, survey cost information lags behind reporting of paradata. Furthermore, in some situations, the measurement of costs at the case level

James Wagner is a Research Professor in the University of Michigan's Survey Research Center (UM SRC), 4053 ISR, 426 Thompson St., Ann Arbor, MI 48104, USA. Lena Centeno is a Senior Study Director, Richard Dulaney is a Vice President, Large Survey Practice, and Brad Edwards is a Vice President and Lead Scientific/Methodology Advisor with Westat, 1600 Research Blvd, Rockville, MD 20850, USA. Z. Tuba Suzer-Gurtekin is an Assistant Research Scientist with the University of Michigan SRC, P.O. Box 1248, Ann Arbor, MI 48106, USA. Stephanie Coffey is a Principal Statistician for Demographic Research in the Center for Economic Studies at the US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA.

Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed the portions of this data product that rely on Census Bureau programs or data for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release (approval ID: CBDRB-FY22-CES014-026).

The National Survey of Family Growth (NSFG) was conducted by the Centers for Disease Control and Prevention's (CDCs) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS or the other funding agencies.

*Address correspondence to James Wagner, University of Michigan SRC, 4053 ISR, 426 Thompson St., Ann Arbor, MI 48104, USA; E-mail: jameswag@umich.edu.

https://doi.org/10.1093/jssam/smad028 Advance access publication 30 August 2023

is difficult. Given the time lag in reporting cost information and the difficulty of assigning costs directly to cases, survey designers and managers have frequently turned to proxy indicators for cost. These proxy measures are often based upon level-of-effort paradata. An example of such a proxy cost indicator is the number of attempts per interview. Unfortunately, little is known about how accurately these proxy indicators actually mirror the true costs of the survey. In this article, we examine a set of these proxy indicators across several surveys with different designs, including different modes of interview. We examine the strength of correlation between these indicators and two different measures of costs—the total project cost and total interviewer hours. This article provides some initial evidence about the quality of these proxies as surrogates for the true costs using data from several different surveys with interviewer-administered modes (telephone, face to face) across three organizations (University of Michigan's Survey Research Center, Westat, US Census Bureau). We find that some indicators (total attempts, total contacts, total completes, sample size) are correlated (average correlation $\sim$0.60) with total costs across several surveys. These same indicators are strongly correlated (average correlation $\sim$0.82) with total interviewer hours. For survey components, three indicators (total attempts, sample size, and total miles) are strongly correlated with both total costs (average correlation $\sim$0.77) and with total interviewer hours (average correlation $\sim$0.86).

KEY WORDS: Monitoring; Paradata; Survey cost.

## Statement of Significance

Survey designers make difficult decisions aimed at maximizing quality and controlling costs. During data collection, aspects of quality are monitored (number of interviews, response rate, differences in the characteristics between those who respond and those who do not). Components of costs are also monitored. Monitoring costs in real time can be difficult. As a result, proxy indicators are often used instead of direct measures of costs. For example, the total number of attempts to complete interviews or the average number of attempts per complete can be more easily calculated in real time than total interviewer hours. The latter may be delayed until timesheets are completed and processed. We explore whether these proxy indicators are good measures of the direct costs of interest.

## 1. INTRODUCTION

Survey design decisions are—by their very nature—tradeoffs between costs and errors. The total survey error framework orients us to produce the highest

quality estimates for a specified budget or, less frequently, produce a specified quality without a prespecified budget. Survey methodologists have frequently studied the error side of this tradeoff. Many studies investigated the impact of nonresponse on the quality of estimates (see Groves 2006; Peytchev 2013 for overviews). There are also studies of the impact of measurement error (Biemer and Trewin 1997; Hox et al. 2015; Fricker et al. 2015; Cernat 2015), sampling error (Kish 1965; Cochran 1977; Lohr 2019), coverage error (Eckman and Kreuter 2011), and even processing error (West et al. 2016, 2017). However, fewer studies examine survey costs (see Olson et al. 2021 for a review).

Further, surveys are growing more complex. Many surveys now include multiple modes (AAPOR Task Force Report—Olson et al. 2021) or responsive and adaptive design elements (Groves and Heeringa 2006; Schouten et al. 2017) that require cost information be available in order to make decisions during data collection. These complexities create new challenges for monitoring and understanding survey costs.

Often, monitoring of survey costs has a longer time lag than monitoring of paradata. Expenses may be reported but need to be routed through financial systems and may not appear in project accounting until a short or even long amount of time has passed. In field surveys, time spent on each case is not usually captured by sample management software. Further, not all Computer-Assisted Telephone Interviewing (CATI) facilities can use the time captured by software for processing interviewer payment. Hence, interviewer hours may only be captured by timesheets. However, timesheets are usually not reported on a daily basis, while records of call attempts are. Further, in some situations, the measurement of costs is difficult (Varela and Zotti 2019; Olson et al. 2021). For example, field interviewers may report hours worked each day, but do not report hours worked on each case. In a field setting, case-level costs are difficult to determine for a cluster sample given that an interviewer travels to the cluster. It is unclear how these travel costs should be divided across all cases in the cluster.

Given the lag between reporting cost information and the difficulty of assigning costs directly to cases, survey designers and managers frequently monitor costs through proxy indicators. These proxy measures are often based upon *level-of-effort* paradata. For example, Groves and Heeringa (2006) use the number of attempts to each sampled unit as a proxy cost indicator in the field studies they describe. The total number of attempts made per complete is another example of a proxy indicator that is based upon paradata. These proxy cost indicators may be reported in experimental comparisons of different survey protocols. The use of these proxy indicators also greatly simplifies the task of monitoring costs on a real-time basis.

Unfortunately, little is known about how accurately these proxy indicators actually mirror the true costs of the survey. This question is an important one for those who monitor surveys during the field period. Understanding whether

these indicators are likely to provide an accurate prediction of true costs is important. Many of these indicators are currently used, without a clear understanding of how and when they work. To make effective decisions, organizations will need to have an evaluation of the ratio of "signal" to "noise" in each of these indicators.

## 2. BACKGROUND

Survey data collections are becoming increasingly varied. More operations include responsive design (Groves and Heeringa 2006; Mohl and Laflamme 2007; Peytchev et al. 2009; Tabuchi et al. 2009; Kleven et al. 2010; Laflamme and Karaganis 2010; Lundquist and Särndal 2013; Barber et al. 2011; Schouten et al. 2011, 2017; Finamore et al. 2013), or more complex designs that include multiple response modes and complex recruitment strategies that may include mail, email, SMS (i.e., text messages), telephone, and face-to-face contacts within the same survey (Olson et al. 2021). These more complex designs can also be tailored to subgroups in the population using adaptive survey design (Schouten et al. 2017). To optimize cost-error tradeoffs for these designs, detailed cost estimates for all the aspects of the design are needed (see for example, Calinescu et al. 2013). However, methods for estimating costs have not been widely discussed in the survey methodology literature. In particular, methods for monitoring costs in real-time have not been evaluated.

Several papers examine total costs (Nicholls and Groves 1986; Catlin and Ingram 1988; Baker 1992; Baker et al. 1995; Cobanoglu et al. 2001; Andresen et al. 2008) or per interview costs for comparison purposes (Kristal et al. 1993; Kaplowitz et al. 2004; Hardigan et al. 2012). Many of these studies report these costs as dollars or as relative costs (Fries et al. 2004; Scott et al. 2011; Biemer et al. 2018), where these costs are evaluated after the survey is complete. However, these studies haven't reported on the correlation of proxy measures with these total costs—whether the costs are expressed as project totals, per interview costs, or relative costs.

On the other hand, proxy indicators have been used as a substitute for more direct cost estimates (Kristal et al. 1993; Gfroerer et al. 2002; Curtin et al. 2005; Cotter et al. 2005; Groves and Heeringa 2006; McCarty et al. 2006; Romanov and Nir 2010). In one of these examples, Gfroerer and colleagues look at "cost savings" resulting from the use of an incentive by examining the reduction in the number of attempts and number of days in the field period required to complete an interview. The survey they examine (the National Survey of Drug Use and Health) is conducted face-to-face and it is presumably difficult to assess the cost of a single attempt. Interviewers must first travel to sampled clusters and then begin making contact attempts. Interviewers were instructed to work at least four hours when traveling to a sampled cluster so as to be efficient.

It is logical to assume that indicators of level of effort would be useful proxy indicators of cost. However, there are reasons to suspect that they may be less than perfect proxies. In a face-to-face survey with a cluster sample, the number of attempts made on each trip to a cluster can vary depending upon how long the interviewer stays, the time of day and day of week, whether they conduct interviews, and many other factors (Wagner and Olson 2018). Estimating the costs of an attempt, in this context, can be difficult. Furthermore, deciding how to assign the cost of travel to a cluster among the cases in that cluster is a difficult problem. Different types of call outcomes, including no one home, ineligible, refusal, and completed interviews, can take different amounts of time, leading the length of individual attempts to vary across sample cases. Interview lengths also vary greatly. Finally, there are reports of errors in the paradata that are used to form these proxy indicators. Biemer et al. (2013) noted that field interviewers underreport attempts. This underreporting reduces the utility of these data for nonresponse adjustments. Wagner et al. (2017a) report on a survey of field interviewers who self-report that they sometimes make mistakes (data entry issues, forgetting to log attempts). Wagner et al. also use GPS data to show the potential for unreported attempts. Therefore, there are reasons to be concerned that these proxy measures may not be highly correlated with the actual costs.

Proxy indicators of costs are being used, both in field monitoring and as an evaluation tool after surveys are completed. However, little is known about the quality of these proxy indicators. In this article, we will examine the quality of these proxy indicators. We do so by examining correlations between several proxy indicators and two measures of costs—total survey costs and total interviewer hours. We also examine the correlation of these proxy cost indicators with costs measured for subcomponents of surveys. The subcomponents that we examine here are interviewers and regions. The analysis we perform provides a roadmap for organizations to conduct similar analyses to more completely understand the quality (i.e., signal-to-noise ratio) of indicators in their own data collection settings. In the next section, we will describe how the indicators are calculated. Then, we will examine correlations between the proxy cost indicators and the two measures of costs. We will conclude with some discussion and recommendations.

## 3. METHODS AND DATA

Table 1 lists the proxy cost indicators used in this study. These indicators were developed from the current monitoring tools in place across several organizations and survey projects. These indicators are used for interviewer-administered surveys (telephone and face to face). Several of these have been reported in the literature as proxy measures of effort. In addition to the proxy indicators and a short description of each, table 1 includes citations that are meant to be examples of the use of each indicator. Some indicators, such as "Total Completes" are reported

**Table 1. Proxy Cost Indicators**

| Indicator | Description | Citations |
|---|---|---|
| Total attempts | A count of the total number of attempts. This includes mail, email, SMS text, telephone, and face-to-face attempts. | Pollien and Joye (2014) and Laflamme (2008) |
| Attempts per interview | Total attempts divided by total completes. | Cotter et al. (2005), Greenlaw and Brown-Welty (2009), Kristal et al. (1993), McCarty et al. (2006), and Romanov and Nir (2010) |
| Total contacts | The count of attempts that have contact with a household member. | |
| Total completes | Total count of completed interviews across all modes. "Complete" is defined separately for each survey. | Uhlig et al. (2014) and Herring et al. (2014) |
| Sample size | The total sample size. This could be a sample of persons or households, depending upon the study. | Vannieuwenhuyze (2013) and Barrett et al. (2017) |
| Attempts per sampled unit | Total attempts divided by the sample size. | See "attempts per interview" citations |
| Total miles | The total number of miles driven by interviewers in their personal vehicles. These miles exclude mileage in rental vehicles. Interviewers using public transportation also do not report mileage. | Wagner and Olson (2018) |
| Mean number of attempts to first contact | For each sampled unit the attempt at which the first contact was recorded. Among those cases with a first contact, what was the mean number of attempts including the attempt when first contact was made. | Luiten et al. (2020), Coffey and Elliott (in press), Laflamme (2008), and Lynn and Clarke (2002) |
| Minutes of interviewing | The total number of minutes spent completing a survey, including all modes except paper/mail surveys. | Adler (1975) and Presser and McCulloch (2011) |

*Continued*

**Table 1.** *Continued*

| Indicator | Description | Citations |
|---|---|---|
| Hours per interview | The total hours interviewers have worked divided by total completes. Hours spent in training are not included. Used as a proxy for total costs only since it includes total interviewer hours in its calculation. | Groves and Heeringa (2006), Kirgis and Lepkowski (2013), and Wagner et al. (2017b) |

by most or even all surveys. "Sample Size" is also commonly reported, although in some cases this would need to be derived from other information, for example, number of completes divided by the response rate.

We compare these proxy indicators to direct measures of costs. For this purpose, we used two measures of survey costs. The first is the total cost of the survey, which is actually a difficult concept to define for a number of reasons (Olson et al. 2021). In some cases, costs may be measured with error. For example, if managers are working many hours more than 40, but can only charge 40 hours, or if managers' salaries are charged to overhead and not directly to projects but they still work some time on projects, then this may lead to mismeasurement of the true effort required. In the reporting of costs, the focus is on the variable costs. The fixed costs are often not included in experimental comparisons. Given these and other difficulties, we describe how total costs are defined in the description of each survey.

Total interviewer hours are a second measure. We exclude hours spent in training. We do not include hours of the immediate supervisors of the interviewers, unless these hours are specifically coded as hours spent in the activity of recruiting sample members or conducting interviews with them. As one would expect, these activity codes are sometimes assigned incorrectly, so some hours may be included or excluded erroneously (Wagner et al. 2017a).

We have initial hypotheses about these proxy indicators:

(1) Among the proxy indicators considered, those that are most directly related to levels of effort (total attempts, total attempts per complete, and total attempts per sampled unit) will have the highest correlations with both measures of cost. Hours per interview (HPI) will have a high correlation with total costs.
(2) Proxy indicators that are indirect levels of effort (sample size, total contacts, total miles, mean number of attempts to first contact, and minutes of interviewing) will have lower correlations than the more direct measures.

Among the proxy indicators of indirect levels, sample size is not directly related to effort as the number of attempts per sampled unit can differ across samples. Total contacts and mean number of attempts to first contact address a component of the recruitment process—contact, but not other parts of the recruitment process—namely obtaining agreement to participate. Total miles are a measure of interviewer travel, but travel in most face-to-face studies is only 20–40 percent of interviewer time (Wagner and Olson 2018). Finally, minutes of interviewing measure how long the actual interview takes, but not other components. While the length of interview may relate to the response rate—longer interviews having lower response rates (most of the research relates to mail and web surveys; Galesic and Bosnjak 2009; Kaplowitz et al. 2012; Rookey et al. 2012; Robb et al. 2017)—the relationship is indirect.

Next, we give a brief description of the surveys that were included as part of this study. In addition to describing the study, we also describe how the survey costs are measured and reported as this varies over the surveys.

## 3.1 National Survey of Family Growth

*3.1.1 Survey description.*
The National Survey of Family Growth (NSFG) is a large face-to-face survey that is conducted nationally. The topic of the survey is fertility and family formation. The target population is women and men ages 15–44 and, since 2015, the eligible ages have been expanded to 15 to 49 (see https://www.cdc.gov/nchs/nsfg/index.htm for additional details on the NSFG, last accessed July 13, 2023). The data analyzed here are from the NSFG 2011–2019 and include data collected between June 2011 and December 2017. The NSFG 2011–2019 had an overall AAPOR RR2 of about 68 percent.

The survey had a continuous design with annual data collection organized into four quarters, each with a new release of sample. Each quarter was 12 weeks long, with four quarters in a year. Although the data were released every two years, each of these quarters could be seen as a repeated cross-sectional survey. We calculate the total costs, interviewer hours, and the proxy cost indicators for each quarter.

*3.1.2 Data description.*
The analyses for the NSFG rely upon cost data as well as proxy indicators that are developed from paradata about the level of effort. The survey paradata are derived from an electronic sample management system and Computer-Assisted Personal Interviewing (CAPI) software. The cost data are drawn from time and expense reporting systems.

Interviewers reported on the outcome of each contact attempt in a sample management system (SurveyTrak). Wagner et al. (2017a) report on a survey of interviewers for this project. Interviewers report that these data are likely

imperfect. While acknowledging these errors, we still use these data to derive the number of completed interviews, the total number of attempts, attempts to first contact, and other indicators that use these counts in their calculation (e.g., mean attempts per completed interview).

Total miles are reported as an expense through time and expense reporting software. Interviewers are asked to report this information at the end of each workday. Interviewers do not always complete this reporting task on the same day. There may be some errors in reporting hours, expenses, or mileage (Wagner et al. 2017a). One issue, discussed in detail by Wagner et al. (2021), is that expenses and even interviewer hours can be classified into the incorrect quarter. These reports are then used as inputs to the University of Michigan (UM) financial systems for payment. The costs reported in this article are based on paid salaries and expenses reported through the UM financial systems. Total miles and total interviewer hours are handled differently. The HPI values are calculated directly from hours reported in the time and expense management software and interview counts reported in the sample management system.

Those central office staff who are paid monthly report their time in a separate timesheet system. Central office staff may work on multiple projects and estimate the hours worked on each project. Staff also categorize their hours into various activities. However, most staff have activity that falls into a single category (e.g., sampling, management). Staff complete their timesheets for a complete month before the end of the month. They estimate hours they expect to incur for the latter part of the month. Any changes in planned hours are then retroactively edited. These hours are uploaded to UM financial systems on a monthly basis and are reported on UM systems approximately ten days after the end of the month.

Minutes of interviewing is taken from the CAPI software (Blaise) and is based on timestamps for entry into the instrument and exit from the instrument (including accounting for possible multiple sessions).

Finally, not all costs are included. We do not include the hours charged by central office management. We do include the hours charged by interviewers for the activity known as "main data collection," which includes the bulk of their hours and expenses. The only interviewer activity excluded is "sampling," which is time charged for development of the sample frame (i.e., listing housing units). We did replicate the analysis with the "sampling" activities included and found very similar results. Time and expenses charged by field supervisors are also included. We included expenses (largely mileage and other travel-related costs) charged by interviewers and respondent payments (tokens of appreciation) in these estimates of costs.

## 3.2 Survey of Consumers

*3.2.1 Survey description.*
The Surveys of Consumers (SoC) are monthly cell phone surveys with a rotating panel design (see https://data.sca.isr.umich.edu/ for additional details on the SoC, last accessed July 13, 2023). Each monthly survey sample is composed of new nationwide Random Digit Dialing (RDD) cell phone sample and also attempted reinterviews of sample previously interviewed 6 and 12 months ago. In a calendar year, the survey period varies across the months. It is 25–27 days in 9 of the months, and 33–34 days for 3 of the months. There are no monetary incentives offered. Reinterview sample cases with a postal or email address are sent an advance notification. In addition, up to two reminder emails are sent in the second half of the survey period. The noncontacts are attempted up to three and ten times in the new and reinterview samples, respectively.

The survey now targets approximately 600 completes each month. At the start of the time period reviewed here (2015), the survey targeted 500 completes. The target total was increased gradually until it reached 600 in November 2016. The survey population is the adult population (18 years and older) located in the coterminous United States (48 states and DC). SoC primarily produces estimates of monthly change in U.S. consumer sentiment and short- and long-term inflation expectations that are used in the nation's monetary and financial policies.

*3.2.2 Data description.*
The average interview length is computed as 29 minutes per complete across all monthly samples from January 2015 to December 2020 using timestamps from the CATI data collection system (Blaise). The average monthly AAPOR RR2 is 5, 52, and 57 percent, respectively, for the fresh, 6-month and 12-month reinterviews. While the telephone numbers are assigned to the interviewers by group and call priorities within the system in each shift in a centralized location using the Blaise sample management system, the interviewers manually call the assigned phone numbers. Each call attempt is recorded by the system at the sample unit level (i.e., telephone number) and the total number of attempts (calls) is computed using the data from the Blaise system. Using the same data, the number of attempts (calls) to first contact is computed using the outcomes defined as "contact" by the study.

The HPI is computed using retrospective reporting in a timesheet by the team leaders and the interviewers of their hours which are associated with a code for each monthly survey. The Blaise system also tracks the time spent interviewing using timestamps. The overall HPI as computed using the interview timestamps from the Blaise data collection system is virtually equal to the average HPI computed using the timesheet hours. The difference could be due to team leaders spending time on interviewing versus administrative work and/or human error.

Our estimate of total costs includes the permanent staff time (programmers, coders, and shift managers) that is tracked by the timesheet system. Some additional functions, for example, human resources, are also included in the total costs. Both staff and interviewers recorded their hours for training separately. Since 2019, costs related to quality control are separated as well.

## 3.3 Survey of Income and Program Participation

*3.3.1 Survey description.*
The Survey of Income and Program Participation (SIPP) is an in-person interviewer-administered national household survey. The survey design is a continuous series of national panels with an annual sample size of about 53,000 households. Data collection operations are conducted annually and last approximately five months (February through June), with the prior calendar year as the reference period (e.g., the 2017 data collection period asks questions about the 2016 calendar year). Each sampled household may be interviewed in four consecutive annual interview periods. Thus, in the full rotating panel implementation, four panels are interviewed simultaneously—the newest panel is being interviewed for the first time (i.e., wave 1), while the oldest panel is being interviewed for the fourth time (i.e., wave 4).

The SIPP sample is a multistage stratified sample of the US civilian noninstitutionalized population. The target population is the civilian noninstitutionalized population of the 50 states and the District of Columbia. In sampled households, all individuals 15 and older are interviewed, and proxy responses are collected for those 14 and younger.

The SIPP is a primary source for information on workforce participation, demographic data, income information, and eligibility for and participation in government assistance programs. The information provided by the SIPP is highly valued by government agencies, academics, and private research organizations. Uses include evaluating the impact of government benefit programs as well as the effects of policy changes to those programs; generating information on the distribution of wealth throughout the country; and informing socioeconomic indicators for the US noninstitutionalized population.

For this analysis, we used SIPP data from 2014 through 2020 which covered the full 2014 panel (2014–2017) and 3 years of the 2018 panel (2018–2020). In both 2014 and 2018, new initial samples were selected, and so those data collection periods only included wave 1 cases. For waves 2–4 in the 2014 panel, respondent cases from wave 1 were recontacted in each year of 2015–2017, but no additional refreshment samples were selected in each of those three later years. As a result, 2015 only included wave 2 cases; 2016 only included wave 3 cases; and 2017 only included wave 4 cases. In 2018, the fully rotating panel design was implemented, so refreshment samples would be selected each year. Figure 1 illustrates these two panel designs. In the 2018

| Year of Entry | Calendar Year of Data Collection | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (Sample Year) | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| 2014 | 1 | 2 | 3 | 4 | | | | |
| 2015 | | | | | | | | |
| 2016 | | | | | | | | |
| 2017 | | | | | | | | |
| 2018 | | | | | 1 | 2 | 3 | 4 |
| 2019 | | | | | | 1 | --- | --- |
| 2020 | | | | | | | 1 | 2 |
| 2021 | | | | | | | | 1 |

**Figure 1. Interview Number (Wave) for SIPP Sample Units by Year of Sampling/ Panel Entry and Calendar Year of Data Collection.**

design, a new wave 1 sample enters each year from 2018 to 2021, with the goal of collecting data for four years. This design change to a rotating panel design was implemented to facilitate cross-sectional as well as longitudinal estimates, improve stability of sample size over time, and address concerns about panel representativeness. Due to extenuating circumstances, the 2019 sample was dropped after wave 1, and so that sample only included wave 1 interviews. In the four months leading up to the start of data collection operations in the SIPP, the federal government (including the Census Bureau) was subject to a continuing resolution in appropriations, a month-long government shutdown, and reductions in overall budgets for governmental agencies. Each of these situations severely impacted preparations for data collection operations, hiring of interviewers, and the budget for data collection activities. This led to significantly lower response rates than those typically achieved in the wave 1 of SIPP.

In figure 2, we report cumulative weighted response rates (AAPOR, RR2) for each panel and wave. For more information on SIPP response rates, see the annual SIPP Source and Accuracy statements (the most recent example is for the 2021 calendar year, Census Bureau 2022).

*3.3.2 Data description.*

Data for these analyses come from five information sources. The sample size and interview/noninterview status come from the US Census Bureau's official field management operational control system (ROSCO). These official statuses are used to estimate response rates, as well as determine how cases are impacted during postcollection processing, including editing, imputation, and weighting.

| Year of Entry | Calendar Year of Data Collection | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (Sample Year) | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| 2014 | 69.8% | 52.3% | 42.2% | 36.9% | | | | |
| 2015 | | | | | | | | |
| 2016 | | | | | | | | |
| 2017 | | | | | | | | |
| 2018 | | | | | 58.4% | 34.3% | 30.6% | 27.9% |
| 2019 | | | | | | 25.4% | --- | --- |
| 2020 | | | | | | | 36.4% | 23.1% |
| 2021 | | | | | | | | 42.5% |

**Figure 2. Cumulative Weighted Response Rates for SIPP Sample Units by Year of Sampling/Panel Entry and Calendar Year of Data Collection.**

Information about the number and outcome of individual contact attempts are extracted from the Contact History Instrument (CHI), within which interviewers self-report each contact attempt, information about the contact attempt, and the outcome of that attempt (e.g., noncontact, contact, refusal, and interview).

Total miles and hours are calculated using information from the time and expense reporting software used by interviewers (WebFRED). Similar to the NSFG, interviewers are asked to input their miles and hours daily, though in practice there may be some lag. However, the assignment of the hours and expenses to the incorrect wave is extremely unlikely since there is a seven-month gap between interview periods, so it is possible to assign interviewing costs to the appropriate interview period. On the other hand, interviewers assigned to work on the SIPP may also be assigned to other surveys as part of their Census Bureau job. When an interviewer makes contact attempts on cases from different surveys on the same day, allocating time and mileage across the different survey projects can be difficult, so there may be errors in reported hours, expenses, or mileage.

Minutes of interviewing is calculated from the CAPI survey software (Blaise). Blaise records each time an interviewer enters and exits the survey instrument for a given case and also provides timestamps to calculate the time spent on any specific screen of the instrument. Aggregating these times results in an estimate of the total time spent interviewing.

We estimated "total" survey costs two ways. For the more restrictive estimate of cost, we limited costs to those incurred by interviewers as a part of their job. This definition of cost includes planning their trips for a day, driving, making contact attempts, and interviewing. This definition does not include any regional office (RO) or Census Bureau headquarters costs, as it focuses

specifically on interviewer activities. We were able to do this using WebFRED data and filtering to only include charges related to interviewers. One limitation of this definition is that interviewing-related costs accrued by noninterviewer staff who may occasionally conduct interviews (interviewer supervisors, trainers, etc.) will not be included in this estimate of "survey costs" as their time is not recorded in the same system. We would expect this definition of cost to be highly correlated with interviewer-level cost proxies (e.g., total miles and minutes interviewing) but less representative of the overall cost of conducting a data collection operation.

We also considered a broader definition of "total" survey costs—the total budget allocations to the field ROs. This estimate of survey costs includes some fixed costs like hiring, training, and survey equipment like laptops, in addition to interviewer-incurred costs, and so it is a more complete way of looking at data collection costs. However, this estimate of cost may be more invariant to specific sample size fluctuations, and more representative of the "cost of doing business" aspect of survey operations. This cost definition still excludes Census Bureau headquarters costs.

While the SIPP has a panel design, all costs and measures of effort for a given data collection period are aggregated across panels. This may not be ideal, but the data are structured such that we cannot disaggregate costs by panel. An interviewer's case load can include cases from all waves, and some costs like mileage and other travel expenses cannot be easily allocated to individual cases as interviewers visit many cases in a given day.

## 3.4 Medical Expenditure Panel Survey

*3.4.1 Survey description.*

The Medical Expenditure Panel Survey (MEPS) is the primary comprehensive source of data on medical services utilization and expenditures and one of the most comprehensive on health insurance coverage for the US noninstitutionalized population. Observing the recruitment of its 26th panel in 2021, the MEPS Household Component (MEPS-HC) captures, over five interviews completed approximately every six months, two calendar years of data about participating households' medical services use, costs, and quality; health conditions and status; and health insurance coverage. Each year a new nationally representative sample for the MEPS-HC is drawn from among households responding to the previous year's National Health Interview Survey (NHIS). One knowledgeable household member (18+) reports for all related household members.

The sample for the MEPS panel 24 was first fielded in 2019. The sample was randomly selected from among those who participated in the NHIS during the first three quarters of 2018. The AAPOR response rate 1 for panel 24 round 1 was 71.2 percent (unweighted conditional response rate) and was calculated by dividing the number of completed household interviews (7,186), by the number

of eligible households (10,090). This rate was conditioned on an NHIS complete or partially completed interview. The final weighted response is not included in this description because it blends all active panels for the calendar year and there were three active panels for 2019. Therefore, the unweighted conditional response is included, which only accounts for the panel MEPS used in this analysis.

*3.4.2 Data description*

For MEPS, the total cost was calculated at the interviewer level using data that were extracted from the time and expense interviewer management system which interviewers use to input their hours and expenses. Only direct costs, such as interviewer wages and expenses were included in this calculation. Hours dedicated to training or indirect costs were not factored into this total cost calculation. Additionally, we excluded field management and home office staff hours.

The number of observations for the Medical Expenditure Panel Survey (one round of data collection over a six-month period) is much smaller than for the other studies in this article. We were concerned that the overall MEPS correlations might be unstable because of the small number. Instead, we examine the proxy indicators and their correlation with interviewer-level hours and costs in a single year. We chose to focus on the MEPS panel 24 and its initial interviews (MEPSP24R1, i.e., MEPS panel 24 round 1; $n \sim 7{,}000$), which occurred between January and July 2019. We estimated correlations using data from 332 interviewers. These interviewers worked on two other panels at the same time as the initial interviews for panel 24 were conducted 24 (MEPS panel 23 round 3, MEPS panel 22 round 5). They charged their time separately for work on the three panels. Though there is certainly error and variation in how individual interviewers allocated time across panels (similar to the issue described above for SIPP interviewers who worked on other projects), the proportion of total survey hours allocated to the initial interviews has been surprisingly consistent for more than 25 years.

We summarize key features of the four surveys in table 2.

## 4. RESULTS

We first look at correlations of proxy indicators with total costs (as defined by each survey). Table 3 shows these correlations for each of the surveys. Here, we note that some indicators are "not applicable" and some are not calculated. "Total miles" is not applicable for the SoC since it does not use face-to-face interviewing. Other indicators are not calculated since it would have been difficult to do so retrospectively. These are "mean number of attempts to first contact" for the SoC; total attempts and sample size for interviewers for the MEPS; and refusal for the ROs for the SIPP.

**Table 2. Key Features of Each Survey**

| | NSFG | SoC | SIPP | MEPS |
|---|---|---|---|---|
| Data collection organization | UM SRC | UM SRC | US Census Bureau | Westat |
| Mode | Face to face | Telephone | Primarily face to face | Face to face |
| Population | 18–49 | 18+ | All members of a sampled household (<14 interviewed by proxy) | 18+, civilian, noninstitutionalized population |
| Design | Cross-sectional | Rotating panel | Mix of longitudinal and rotating panel | Rotating panel |
| Unit of analysis | Quarterly samples | Monthly samples | Annual samples/regional offices within annual samples | Interviewers within a single round of interviewing within a panel |
| Number of units for analysis | 27 quarters | 72 months | 6 regional offices, 7 years | 332 interviewers |
| Average sample size | ∼5,000 housing units | ∼8,600 cell numbers | 53,000 housing units | 9,700 new households were sampled from NHIS |
| Sample source | Area probability samples | Cellular RDD | Master address file | NHIS completed or partially completed interviews |
| Response rate | ∼68% | 5% panel, conditional 52% 6-month follow-up, conditional 57% 12-month follow-up | See figure 2 | ∼70% |

*Continued*

**Table 2.** *Continued*

| | NSFG | SoC | SIPP | MEPS |
|---|---|---|---|---|
| Length of sample period | 12 weeks | 1 month | 5 months | 30 months |
| Contact strategies | Personal visits, mailings | Telephone | Personal visits, telephone contacts by field interviewers | Personal visits, telephone contacts by field interviewers, advance mailing by home office |
| Cost definitions | Total cost calculated as interviewer wages and expenses with some exclusions such as training costs and costs related to sample development. | Total cost calculated as interviewer and permanent staff time wages and expenses that are directly related to the data collection. Permanent staff includes programmers, coders, and shift managers. There are some additional functions such as human resources support as well. | Interviewer costs: Interviewing-related costs only (e.g., planning work, traveling to cases, making contact attempts, interviewing activities) Interviewer + RO costs: Total budget allocations to field activities (e.g., hiring, training, equipment, in addition to interviewer-specific costs) | Total cost was calculated at the interviewer level using data from time and expense interviewer management system. Only direct costs were included, that is, interviewer wages and expenses. |

**Table 3. Correlations of Proxy Indicators with Total Cost Estimate**

| Survey (# of iterations) | Proxy indicator | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total attempts | Attempts per interview | Attempts per sampled line | Total contacts | Total completes | Sample size | Total miles | Mean number of attempts to first contact | Minutes of interviewing | Hours per interview |
| NSFG | 0.496[b] | 0.266 | 0.409 | 0.370 | 0.309 | 0.319 | 0.203 | 0.156 | 0.218 | 0.509[a] |
| SoC | 0.529 | 0.341 | −0.361 | 0.497 | 0.817 | 0.826[a] | NA | NC | 0.222 | 0.735[b] |
| SIPP v1 I'wer costs | 0.595 | 0.333 | 0.012 | 0.706 | 0.430 | 0.674 | 0.804[b] | 0.175 | 0.818[a] | 0.688 |
| SIPP v2 I'wer + RO Costs | 0.787 | 0.072 | 0.645 | 0.808 | 0.916[a] | 0.542 | 0.454 | 0.880[b] | 0.772 | −0.262 |

NOTE.—NA: not applicable; NC: not calculated.
[a]Highest correlation per study.
[b]Second highest correlation per study.

In this table, we see some similarities and differences across the correlations. Nearly all the correlations are positive. The correlations tend to be stronger for the SIPP study using the second version of total costs (i.e., those including both interviewer and RO costs) than for the other studies, though this may be because of the small number of years on which correlations are based. The indicator with the highest average correlation is the "total completes." There are several indicators that are close to this, including "total attempts," "total contacts," and "sample size." Among the remaining indicators, several have average correlations between 0.4 and 0.5 ("total miles," "mean number of attempts to first contact," "minutes of interviewing," and "hours per interview"). Two indicators have low average correlations ("attempts per interview" and "attempts per sampled unit").

Next, we look at the correlations of these proxy indicators with total interviewer hours. Table 4 presents these correlations.

In this table, the correlations are generally stronger. The highest correlation for the NSFG is the total number of contacts. Total attempts is a close second highest correlation for the NSFG. For the SIPP the highest correlations are the total contacts and the minutes of interviewing. For the SoC, the sample size has the highest correlation with total interviewer hours. Overall, total attempts and total contacts have the highest average correlation across the three studies. Total attempts has the highest minimum correlation among all three studies.

We next turn our attention to correlations within two of the studies, MEPS and SIPP. Here, we generically refer to subsets of surveys as "components." In the MEPS, the component is the interviewer. In the case of the SIPP, the component is the RO (the US Census Bureau has six ROs). For the MEPS we look at correlations of interviewer-level total costs and each of the proxy indicators, also calculated at the interviewer level. For the SIPP, we look at correlations of RO-level costs with proxy indicators calculated at the same level, using the two different cost calculation approaches. The results are presented in table 5. These correlations help us answer the question of whether these proxy indicators are useful at levels below the survey.

For the MEPS, the total miles is the proxy indicator with the highest correlation with costs. Total miles and minutes of interviewing have high correlations with total costs for the SIPP. Some of the indicators show relatively weak correlations with total costs, including attempts per interviewer and attempts per sample line. The indicator with the highest average correlation is total attempts (only observed for SIPP). Among indicators reported for all three rows, total miles has the highest correlation. The highest minimum correlation is for total attempts (based on the two SIPP rows) and sample size (based on all three rows).

Table 6 looks at the correlations of the proxy indicators with total interviewer hours, again at the component level.

In this case, there are some negative correlations. For example, higher attempts per interview are associated with lower interviewer hours (see section

**Table 4. Correlations of Proxy Indicators with Total Interviewer Hours**

| Survey | Proxy indicator | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total attempts | Attempts per interview | Attempts per sampled line | Total contacts | Total completes | Sample size | Total miles | Mean number of attempts to first contact | Minutes of interviewing |
| NSFG | 0.857[b] | 0.183 | 0.566 | 0.861[a] | 0.808 | 0.699 | 0.713 | 0.141 | 0.705 |
| SIPP | 0.877 | 0.378 | 0.315 | 0.953[b] | 0.752 | 0.842 | 0.721 | 0.441 | 0.963[a] |
| SoC | 0.815[b] | 0.666 | −0.071 | 0.719 | 0.786 | 0.929[a] | NA | −0.103 | 0.232 |

NOTE—NA: not applicable.
[a]Highest correlation per study.
[b]Second highest correlation per study.

**Table 5. Correlations of Proxy Indicators with Total Cost Estimate for Each Component**

| Survey | Proxy indicator | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total attempts | Attempts per interview | Attempts per sampled line | Total contacts | Total completes | Sample size | Total miles | Mean number of attempts to first contact | Minutes of interviewing | Hours per interview | Refusals |
| MEPSP24R1 Component=I'wers | NC | −0.271 | −0.055 | 0.050 | 0.229 | NC | 0.751[b] | 0.053 | 0.469[a] | 0.387 | −0.217 |
| SIPP Regional Offices, I'wer Costs Only | 0.756 | −0.062 | −0.125 | 0.837 | 0.728 | 0.825 | 0.872[a] | 0.031 | 0.893[b] | 0.344 | NC |
| SIPP Regional Offices, I'wer + RO Costs | 0.805 | −0.149 | 0.291 | 0.812 | 0.860[b] | 0.701 | 0.638 | 0.508 | 0.831[a] | −0.150 | NC |

NOTE.—NC: not calculated.
[a]Highest correlation per study.
[b]Second highest correlation per study.

**Table 6. Correlations of Proxy Indicators with Total Interviewer Hours for Each Component**

| Survey | Proxy indicator | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total attempts | Attempts per interview | Attempts per sampled line | Total contacts | Total completes | Sample size | Total miles | Mean number of attempts to first contact | Minutes of interviewing | Refusals |
| MEPSP24R1 component = I'wers ($n = 332$) | NC | −0.305 | −0.055 | 0.033 | 0.247 | NC | 0.782[b] | 0.069 | 0.504[a] | −0.265 |
| SIPP regional offices ($n = 42$) | 0.878 | −0.087 | 0.005 | 0.933[a] | 0.865 | 0.899 | 0.827 | 0.143 | 0.963[b] | NC |

NOTE.—NC: not calculated.
[a]Highest correlation per study.
[b]Second highest correlation per study.

5 for explanation). The highest correlations are total miles and minutes of interviewing. Total miles has the highest average correlation for indicators with at least two observations. Total contacts also has a high correlation with interviewer hours at the RO level for the SIPP.

## 5. DISCUSSION

The proxy cost indicators tested in this article do seem to be at least somewhat correlated and, in many cases, strongly correlated with both total survey costs and total interviewer hours. We find that some of these indicators (total attempts, total contacts, total completes, sample size) are somewhat correlated (average correlation $\sim$0.60) with total costs across several surveys. These same indicators are strongly correlated (average correlation $\sim$0.82) with total interviewer hours. For survey components, three indicators (total attempts, sample size, and total miles) are strongly correlated with both total costs (average correlation $\sim$0.77) and total interviewer hours (average correlation $\sim$0.86). For three of the studies (i.e., those other than the second version of SIPP costs that includes RO costs), the HPI measure—which includes total interviewing hours as part of its calculation—was highly correlated with total costs.

Total interviewer hours was highly correlated with total attempts for all three surveys examined. The indicator with the highest correlation varied across the three surveys. In general, the proxy indicators were much more highly correlated with interviewer hours than overall costs. This is useful information for surveys and is likely driven by the fact that most proxies are based on fieldwork and activities conducted by interviewers. Although interviewer hours are often an important component of survey costs, there are other components.

When we look at correlations with costs of survey components, we find a more uneven pattern. For the interviewer subcomponent in the MEPS, the correlations are sometimes negative and sometimes positive. Furthermore, some of the correlations are small, while some are large for both total costs and interviewer hours. For both categories of costs, the total miles has the largest correlation and minutes of interviewing is the next largest correlation. Attempts per interview and refusals are both negatively correlated with costs—that is, interviewers with fewer attempts per interview and fewer refusals had higher costs. This may be a result of interviewers who are efficient and who do better than average in gaining cooperation are assigned more sample by managers. This leads to them having higher costs. In the SoC, the attempts per sampled unit have a negative correlation with costs. In part, this can be explained by the survey design that fixes the number of attempts per sampled unit. Further, there were two changes in the design over the field period that largely explain this negative correlation. First, starting in 2015, the number of interviews was increased. As a result, the total costs tended to rise over the following months

even though the attempts per sampled unit were fixed. Second, after March 2020, the efficiency of the survey increased with higher contact and interview rates. The SoC interviewed more cases than usual during this time period. In the SIPP, we find that correlations between proxies and interviewer costs are higher when examining these correlations by RO (table 5) than overall (table 3), suggesting that correlations at component levels of the SIPP may be useful, as ROs may accrue costs differently. For example, interviewers in more rural regions may have to drive farther to attempt cases than interviewers in urban areas, leading to increased mileage costs. Also on the SIPP, there is a negative correlation between HPI and total costs for the costs including RO costs. HPI is only the interviewer hours. The HPI and total interviewer costs have a strong, positive correlation (0.69). The total costs including the RO costs add in the costs of all activities at the RO associated with a survey. This reduces the correlation because so many more activities go into those overall costs beyond interviewer hours. Things like total attempts, total contacts, and total completes still are still correlated with total RO costs because data collection is a large part of overall costs, but when we start dividing one proxy measure by another (e.g., hours per complete), the relationship is less clear.

In general, the proxy indicators explored in this article contain useful information. The number of total attempts appears to be a strong proxy for interviewer hours and, given that interviewer hours are a large component of the total costs for each of these surveys, also with total costs. Other indicators vary in their utility across the surveys and components. Although total contacts, total completes, and sample size were correlated with costs at the survey level, these indicators were less useful for the components explored here. Furthermore, sample size is not useful for monitoring, unless the sample size changes during data collection. It may be useful for organizations using these indicators to evaluate these correlations in order to understand the relative signal-to-noise ratio that may be specific to their setting.

Some of these indicators may also provide important signals that are indirectly related to costs. For example, mean calls to first contact may provide an indication of whether the process of establishing contact with sampled units is being carried out as efficiently as expected. Mileage may also be an indication of issues with contact or recruitment (Wagner and Olson 2018).

Finally, using multiple indicators may be useful for providing early signals when cost issues are developing. These noisy signals may result in "false positives," but further detailed analysis may determine whether there is an actual issue. These additional analyses—which may be expensive to conduct—can look across the indicators explored here as well as others not explored. Statistical models may also be used to predict costs (Wagner 2019; Wagner et al. 2020). In this way, survey designers can make informed decisions about tradeoffs involving costs during data collection. Of course, the error implications need to be considered as well. Here, there are methods developed for monitoring sampling error (Wagner et al. 2012), nonresponse error

(Vandenplas et al 2017; Moore et al. 2018; Coffey et al. 2020), and measurement error (Schouten and Calinescu 2013).

There are some limitations to this study. First, each of the studies only looked at variable field costs. Fixed costs were ignored. In fact, not all variable costs were included, such as interviewer training, costs for equipment like laptops, or maintenance of the interviewing instrument or software. Furthermore, each study defined total costs in different ways. However, this is not an easy issue to address as different data collection agencies and even different surveys within the same organization may have different approaches to tracking and reporting costs. Aligning these approaches across surveys and organizations is a laudable, though likely difficult to achieve goal. Instead, we have tried to be transparent in our descriptions of how these costs are calculated. This should aid in translating findings from the surveys described in this article to other surveys in other settings.

We also acknowledge that there are measurement errors in both paradata and costs. Some interviewer effort (attempts) might not be recorded. Mileage might be incorrectly tracked by interviewers. The scale of these types of measurement error is largely unknown. These measurement errors would likely lead to reduced correlations. Measurement issues may play a role in why some indicators have stronger correlations with costs or hours than others. Learning about these measurement issues should help with interpreting the meaning of proxy indicators and could lead to reductions in these same errors.

There are a number of areas for future research. First, if some of these proxy cost indicators are useful in some contexts but not others, why? What is the difference between these situations? To answer this question, a larger number of studies need to be included and important characteristics of each "coded." Second, could multivariate indicators more complex than the ratios included in this study create better proxy cost indicators? One could imagine estimating a regression model with the dependent variable being costs (as hours, total costs, or something else) and the predictors including the indicators included in this study and then using the estimated coefficients to create a predicted value that is more highly correlated with the dependent cost variable. The predicted value could then be used to monitor costs.

Finally, since relating costs and errors should be the final goal, work to establish relationships between these proxy indicators and error sources—especially nonresponse error, but others as well (i.e., sampling)—will be an important next step. Survey costs are an understudied area. Increasing our understanding of the strengths and weaknesses of these proxy indicators is a promising area of research. One reason for this is that it does not require researchers to disclose detailed—possibly proprietary—financial information. These indicators are already being reported in the literature as findings from experiments. Understanding how they generalize across surveys and organizations will be an important step in allowing research on cost-error tradeoffs to be translated into practice.

## REFERENCES


Adler, L. (1975), "How to Economize on Industrial Marketing Research," *Industrial Marketing Management*, 4, 243–247.

Andresen, E. M., Machuga, C. R., Van Booven, M. E., Egel, J., Chibnall, J. T., and Tait, R. C. (2008), "Effects and Costs of Tracing Strategies on Nonresponse Bias in a Survey of Workers with Low-Back Injury," *Public Opinion Quarterly*, 72, 40–54.

Baker, R. P. (1992), " New Technology in Survey Research: Computer-Assisted Personal Interviewing (CAPI)," *Social Science Computer Review*, 10, 145–157.

Baker, R. P., Bradburn, N. M., and Johnson, A. (1995), "Computer-Assisted Personal Interviewing: An Experimental Evaluation of Data Quality and Costs," *Journal of Official Statistics*, 11, 415–434.

Barber, J. S., Kusunoki, Y., and Gatny, H. H. (2011), "Design and Implementation of an Online Weekly Survey to Study Unintended Pregnancies: Preliminary Results," *Vienna Yearbook of Population Research*, 9, 327–334.

Barrett, B. N., van Poorten, B., Cooper, A. B., and Haider, W. (2017), "Concurrently Assessing Survey Mode and Sample Size in Off-Site Angler Surveys," *North American Journal of Fisheries Management*, 37, 756–767.

Biemer, P. P., Chen, P., and Wang, K. (2013), "Using Level-of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176, 147–168.

Biemer, P. P., Murphy, J., Zimmer, S., Berry, C., Deng, G., and Lewis, K. (2018), "Using Bonus Monetary Incentives to Encourage Web Response in Mixed-Mode Household Surveys," *Journal of Survey Statistics and Methodology*, 6, 240–261.

Biemer, P. P., and Trewin, D. (1997), "A Review of Measurement Error Effects on the Analysis of Survey Data," in *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin. New York: Wiley, pp. 601–632.

Calinescu, M., Bhulai, S., and Schouten, B. (2013), "Optimal Resource Allocation in Survey Designs," *European Journal of Operational Research*, 226, 115–121.

Catlin, G., and Ingram, S. (1988), "The Effects of Cati on Costs and Data Quality: A Comparison of Cati and Paper Methods in Centralized Interviewing," in *Telephone Survey Methodology*, eds. R. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, and J. Waksberg, New York: Wiley, pp. 437–450.

Census Bureau (2022), "Survey of Income and Program Participation: Source and Accuracy Statement. Technical Report," Available at https://www.census.gov/programs-surveys/sipp/tech-documentation/source-accuracy-statements.html. Accessed July 13, 2023.

Cernat, A. (2015), "Impact of Mode Design on Measurement Errors and Estimates of Individual Change," *Survey Research Methods*, 9, 83–99.

Cobanoglu, C., Warde, B., and Moreo, P. J. (2001), "A Comparison of Mail, Fax and Web-Based Survey Methods," *International Journal of Market Research*, 43, 1–452.

Cochran, W. G. (1977), *Sampling Techniques*, New York: Wiley.

Coffey, S., and Elliott, M. R. (in press), "Predicting Days to Respondent Contact in Cross-Sectional Surveys Using a Bayesian Approach," *Journal of Official Statistics*.

Coffey, S., Reist, B., and Miller, P. V. (2020), "Interventions on-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates," *Journal of Survey Statistics and Methodology*, 8, 726–747.

Cotter, R. B., Burke, J. D., Stouthamer-Loeber, M., and Loeber, R. (2005), "Contacting Participants for Follow-Up: How Much Effort is Required to Retain Participants in Longitudinal Studies?," *Evaluation and Program Planning*, 28, 15–21.

Curtin, R., Presser, S., and Singer, E. (2005), "Changes in Telephone Survey Nonresponse over the past Quarter Century," *Public Opinion Quarterly*, 69, 87–98.

Eckman, S., and Kreuter, F. (2011), "Confirmation Bias in Housing Unit Listing," *Public Opinion Quarterly*, 75, 139–150.

Finamore, J., Coffey, S., and Reist, B. (2013), "National Survey of College Graduates: A Practice-Based Investigation of Adaptive Design," Paper presented at the Annual Conference of the American Association for Public Opinion, Boston, MA.

Fricker, S., Kopp, B., Tan, L., and Tourangeau, R. (2015), "A Review of Measurement Error Assessment in a U.S. Household Consumer Expenditure Survey," *Journal of Survey Statistics and Methodology*, 3, 67–88.

Fries, B. E., James, M., Hammer, S. S., Shugarman, L. R., and Morris, J. N. (2004), "Is Telephone Screening Feasible? Accuracy and Cost-Effectiveness of Identifying People Medically Eligible for Home-and Community-Based Services," *The Gerontologist*, 44, 680–688.

Galesic, M., and Bosnjak, M. (2009), "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey," *The Public Opinion Quarterly*, 73, 349–360.

Gfroerer, J. C., Eyerman, J., and Chromy, J. R. (2002), *Redesigning an Ongoing National Household Survey: Methodological Issues*, Department of Health and Human Services Publication No. SMA 03-3768. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Greenlaw, C., and Brown-Welty, S. (2009), "A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost," *Evaluation Review*, 33, 464–480.

Groves, R. M. (2006), "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70, 646–675.

Groves, R. M., and Heeringa, S. G. (2006), "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 439–457.

Hardigan, P. C., Succar, C. T., and Fleisher, J. M. (2012), "An Analysis of Response Rate and Economic Costs between Mail and Web-Based Surveys among Practicing Dentists: A Randomized Trial," *Journal of Community Health*, 37, 383–394.

Herring, P., Butler, T., Hall, S., Bennett, H., Montgomery, S. B., and Fraser, G. (2014), "Recruiting and Motivating Black Subjects to Complete a Lengthy Survey in a Large Cohort Study: An Exploration of Different Strategies," *BMC Medical Research Methodology*, 14, 46.

Hox, J. J., De Leeuw, E. D., and Zijlmans, E. A. O. (2015), "Measurement Equivalence in Mixed Mode Surveys," *Frontiers in Psychology*, 6, 87.

Kaplowitz, M. D., Hadlock, T. D., and Levine, R. (2004), "A Comparison of Web and Mail Survey Response Rates," *Public Opinion Quarterly*, 68, 94–101.

Kaplowitz, M. D., Lupi, F., Couper, M. P., and Thorp, L. (2012), "The Effect of Invitation Design on Web Survey Response Rates," *Social Science Computer Review*, 30, 339–349.

Kirgis, N., and Lepkowski, J. (2013), "Design and Management Strategies for Paradata-Driven Responsive Design: Illustrations from the 2006-2010 National Survey of Family Growth," in *Improving Surveys with Paradata: Analytic Uses of Process Information*, ed. F. Kreuter. Hoboken, NJ: Wiley, pp. 121–144.

Kish, L. (1965), *Survey Sampling*, New York: J. Wiley.

Kleven, Ø., Fosen, J., Lagerstrøm, B., and Zhang, L.-C. (2010), "The Use of R-Indicators in Responsive Survey Design–Some Norwegian Experiences," Paper presented at the European Conference on Quality in Official Statistics, Helsinki, Finland.

Kristal, A. R., White, E., Davis, J. R., Corycell, G., Raghunathan, T., Kinne, S., and Lin, T.-K. (1993), "Effects of Enhanced Calling Efforts on Response Rates, Estimates of Health Behavior, and Costs in a Telephone Health Survey Using Random-Digit Dialing," *Public Health Reports*, 108, 372.

Laflamme, F. (2008), "Understanding Survey Data Collection through the Analysis of Paradata at Statistics Canada," Paper presented at the Annual Conference of the American Association for Public Opinion Research, New Orleans, LA.

Laflamme, F., and Karaganis, M. (2010), "Implementation of Responsive Collection Design for Cati Surveys at Statistics Canada," Paper presented at the European Conference on Quality in Official Statistics, Helsinki, Finland.

Lohr, S. L. (2019), *Sampling: Design and Analysis*. Boca Raton, FL: CRC Press.

Luiten, A., Hox, J., and de Leeuw, E. (2020), "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys," *Journal of Official Statistics*, 36, 469–487.

Lundquist, P., and Särndal, C.-E. (2013), "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey," *Journal of Official Statistics*, 29, 557–582.

Lynn, P., and Clarke, P. (2002), "Separating Refusal Bias and Non-Contact Bias: Evidence from Uk National Surveys," *Journal of the Royal Statistical Society: Series D (the Statistician)*, 51, 319–333.

McCarty, C., House, M., Harman, J., and Richards, S. (2006), "Effort in Phone Survey Response Rates: The Effects of Vendor and Client-Controlled Factors," *Field Methods*, 18, 172–188.

Mohl, C., and Laflamme, F. (2007), "Research and Responsive Design Options for Survey Data Collection at Statistics Canada," Paper presented at the Joint Statistical Meetings, Salt Lake City, UT.

Moore, J. C., Durrant, G. B., and Smith, P. W. F. (2018), "Data Set Representativeness during Data Collection in Three Uk Social Surveys: Generalizability and the Effects of Auxiliary Covariate Choice," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 229–248.

Nicholls II, L., and Groves, R. M. (1986), "The Status of Computer-Assisted Telephone Interviewing: Part I-Introduction and Impact on Cost and Timeliness of Survey Data," *Journal of Official Statistics*, 2, 93.

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., and 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, 381–411.

Olson, K., Wagner, J., and Anderson, R. (2021), "Survey Costs: Where Are We and What is the Way Forward?" *Journal of Survey Statistics and Methodology*, 9, 921–942.

Peytchev, A. (2013), "Consequences of Survey Nonresponse," *The Annals of the American Academy of Political and Social Science*, 645, 88–111.

Peytchev, A., Baxter, R. K., and Carley-Baxter, L. R. (2009), "Not All Survey Effort Is Equal: Reduction of Nonresponse Bias and Nonresponse Error," *Public Opinion Quarterly*, 73, 785–806.

Pollien, A., and Joye, D. (2014), "Patterns of Contact Attempts in Surveys," in *Advances in Sequence Analysis: Theory, Method, Applications*, eds. P. Blanchard, F. Bühlmann and J.-A. Gauthier. Cham, Springer International Publishing, pp. 285–304.

Presser, S., and McCulloch, S. (2011), "The Growth of Survey Research in the United States: Government-Sponsored Surveys, 1984-2004," *Social Science Research*, 40, 1019–1024.

Robb, K. A., Gatting, L., and Wardle, J. (2017), "What Impact Do Questionnaire Length and Monetary Incentives Have on Mailed Health Psychology Survey Response?" *British Journal of Health Psychology*, 22, 671–685.

Romanov, D., and Nir, M. (2010), "Get It or Drop It? Cost-Benefit Analysis of Attempts to Interview in Household Surveys," *Journal of Official Statistics*, 26, 165–191.

Rookey, B. D., Le, L., Littlejohn, M., and Dillman, D. A. (2012), "Understanding the Resilience of Mail-Back Survey Methods: An Analysis of 20 Years of Change in Response Rates to National Park Surveys," *Social Science Research*, 41, 1404–1414.

Schouten, B., Peytchev, A., and Wagner, J. (2017), *Adaptive Survey Design*, CRC Press.

Schouten, B., and Calinescu, M. (2013), "Paradata as Input to Monitoring Representativeness and Measurement Profiles: A Case Study of the Dutch Labour Force Survey," in *Improving Surveys with Paradata*, ed. F. Kreuter, New York, NY: Wiley, pp. 231–258.

Schouten, B., Shlomo, N., and Skinner, C. (2011), "Indicators for Monitoring and Improving Representativeness of Response," *Journal of Official Statistics*, 27, 231–253.

Scott, A., Jeon, S.-H., Joyce, C. M., Humphreys, J. S., Kalb, G., Witt, J., and Leahy, A. (2011), "A Randomised Trial and Economic Evaluation of the Effect of Response Mode on Response Rate, Response Bias, and Item Non-Response in a Survey of Doctors," *Bmc Medical Research Methodology*, 11, 126.

Tabuchi, T., Laflamme, F., Phillips, O., Karaganis, M., and Villeneuve, A. (2009), "Responsive Design for the Survey of Labour and Income Dynamics," in Proceedings of the Statistics Canada Symposium.

Uhlig, C. E., Seitz, B., Eter, N., Promesberger, J., and Busse, H. (2014), "Efficiencies of Internet-Based Digital and Paper-Based Scientific Surveys and the Estimated Costs and Time for Different-Sized Cohorts," *PLoS One*, 9, e108441.

Vandenplas, C., Loosveldt, G., and Beullens, K. (2017), "Fieldwork Monitoring for the European Social Survey: An Illustration with Belgium and the Czech Republic in Round 7," *Journal of Official Statistics*, 33, 659–686.

Vannieuwenhuyze, J. (2013), "On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys," *Survey Research Methods*, 8, 31–42.

Varela, K., and Zotti, A. (2019), "The Complications of Building a Cost Estimate for a National, Multiphase Survey," Paper presented at the 6th International Workshop on Advances in Adaptive and Responsive Survey Designs, Washington, DC.

Wagner, J., West, B. T., Elliott, M. R., and Coffey, S. (2020), "Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context," *Journal of Official Statistics*, 36, 907–931.

Wagner, J. (2019), "Estimation of Survey Cost Parameters Using Paradata," *Survey Practice*, 12, 1–10.

Wagner, J., Guyer, H., and Evanchek, C. (2021), "Using Time Series Models to Understand Survey Costs," *Journal of Survey Statistics and Methodology*, 9, 943–960.

Wagner, J., and Olson, K. (2018), "An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys," *Journal of Official Statistics*, 34, 211–237.

Wagner, J., Olson, K., and Edgar, M. (2017a), "The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata," *Survey Research Methods*, 11, 219–233.

Wagner, J., West, B. T., Guyer, H., Burton, P., Kelley, J., Couper, M. P., and Mosher, W. D. (2017b), "The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth," in *Total Survey Error in Practice*, eds. P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West, New York: Wiley, pp. 155–178.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., and Ndiaye, S. K. (2012), "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection," *Journal of Official Statistics*, 28, 477–499.

West, B. T., Sakshaug, J. W., and Aurelien, G. A. S. (2016), "How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data?," *PloS One*, 11, e0158120.

West, B. T., Sakshaug, J. W., and Kim, Y. (2017), "Analytic Error as an Important Component of Total Survey Error," In *Total Survey Error in Practice*, eds. P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West, New York: John Wiley & Sons. Ltd, pp. 487–510.