DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 625–648, http://dx.doi.org/10.2478/JOS-2018-0030

Transitioning a Survey to Self-Administration using Adaptive, Responsive, and Tailored (ART) Design Principles and Data Visualization

Joe Murphy[1], *Paul Biemer*[2], *and Chip Berry*[3]

This article discusses the critical and complex design decisions associated with transitioning an interviewer-administered survey to a self-administered, postal, web/paper survey. Our approach embeds adaptive, responsive, and tailored (ART) design principles and data visualization during a multi-phased data collection operation to project the outcomes of each phase in preparation for subsequent phases. This requires rapid decision making based upon experimental results using a data visualization system to monitor critical-to-quality (CTQ) metrics and facilitate projections of outcomes from the current phase of data collection to inform the design of the subsequent phase. We describe the objectives of the overall design, the features designed to address these objectives, components of the visual adaptive total design (ATD) system for monitoring quality components and relative costs in real time, and examples of the visualization elements and functionalities that were used in one case study. We also discuss subsequent initiatives to develop an interactive version of the monitoring tool and applications for other studies, including those employing adaptive, responsive, and tailored (ART) designs. Our case study is a series of pilot studies conducted for the Residential Energy Consumption Survey (RECS), sponsored by the U.S. Energy Information Administration (EIA).

Key words: Responsive design, adaptive design, monitoring, data collection, visualization.

1. Introduction

Interviewer-administered survey modes, such as face-to-face and telephone, have traditionally been viewed as the standard for collecting high quality data on nonsensitive topics. The presence of an interviewer can help foster cooperation with the respondent and the interaction between interviewer and respondent can help assure that key questions are interpreted and answered correctly. While they may set the standard for quality, interviewer-administered modes are typically more costly than modes that do not involve interviewers, such as web and paper surveys. Costs for recruiting and training interviewers, their salaries, and their transportation costs (for face-to-face surveys) are major investments for a survey project. From a total survey error perspective, allocating such a large share of the survey budget to face-to-face interviewing may be suboptimal for some

[1] RTI International, 230 W Monroe St. Suite 2100, Chicago, IL, U.S.A, 60606. Email: jmurphy@rti.org
[2] RTI International, 3040 East Cornwallis Road, Post Office Box 12194, Research Triangle Park, NC 27709-2194 U.S.A. Email: ppb@rti.org
[3] U.S. Energy Information Administration; 1000 Independence Ave., SW, Washington, DC 20585, U.S.A. Email: james.berry@eia.gov

surveys (Groves 1989; Biemer 2010). For example, using a less expensive data collection mode could allow a larger sample size, more extensive nonresponse follow-up, more questionnaire pretesting, and the elimination of interviewer error. In addition, using a mail delivery mode obviates the need for cluster sampling that is often required in face-to-face surveys to reduce interviewer travel costs.

In recent years, it has only become more difficult to efficiently collect data, regardless of mode. Response rates have continued to decline and, to achieve acceptable response rates, costs have increased. Some well-established surveys in the U.S. that have employed interviewers in the past such as the Longitudinal Survey of Adolescent Health (Biemer et al. 2017a), the Behavior Risk Factor Surveillance Survey (Link and Mokdad 2005), the National Health Care Interview Survey (Howden et al. 2015) and the Residential Energy Consumption Survey (RECS) (Eddy and Marton 2012) have investigated or are now considering changes to incorporate self-administered designs for cost reduction. However, the change to self-administration is not simple or straightforward when comparable, high quality data are desired using these less expensive modes.

For surveys that are conducted periodically, such as repeated cross-sectional or longitudinal studies, the decision to change modes is only the first in a series of design decisions that must be made before implementing a specific self-administered data collection protocol. To inform this rather complex and challenging transition, a series of pilot studies can be designed to experimentally test a range of alternative designs and identify the best data collection approaches for self-administration, including web and paper modes. Such a design may require that a series of experiments be conducted within a highly compressed schedule with little time between studies for data analysis. Yet, it is essential that the results and lessons learned from each experiment be thoroughly understood and transferred across experiments. As such, design decisions need to be made for the next set of experiments before data collection and analysis for the previous set of experiments are fully completed. Key to decision making is a system for monitoring and visualizing the results of data collection while the survey is in progress. Groves and Heeringa (2006) discuss this problem in the context of a three-phase responsive design for the National Survey of Family Growth where the first phase constituted an experiment that was followed immediately by the main data collection phase. That phase was then followed immediately by a nonresponse follow-up phase. Decisions based upon incomplete data were made during each phase that affected the design of the subsequent phase. To facilitate this process, the study team reviewed daily updates on the results for each treatment or design feature being monitored. This approach involved defining multiple quality components and their metrics as well as a system to compile a vast amount of information for quick, clear, presentation and a minimum of burden on the survey managers.

In this article, we discuss the key design decisions required for the transition of an interviewer-administered survey to self-administration via paper and web using a series of data collection phases. Our approach embeds adaptive, responsive, and tailored (ART) design principles and data visualization during a multi-phased data collection operation to project the outcomes of each phase in preparation for subsequent phases. Key to this process is identifying critical-to-quality (CTQ) metrics to monitor and a data visualization system to meet the requirements of the research strategy. We describe the objectives of the design and system, the features required to address these objectives, and the implementation of a visualization approach for monitoring costs and quality components in real time. We also describe and provide examples of the visualization elements we created and applied as well as their functionalities. Looking forward, we discuss current initiatives to develop an interactive version of the monitoring tool with applications in subsequent studies especially those employing ART designs.

Our case study is a series of pilot studies conducted for the U.S. Energy Information Administration (EIA) for the RECS. The goal of these studies was to assess the operational feasibility, data quality and costs of converting the RECS to a web and paper mixed mode design. The RECS has traditionally been conducted by face-to-face interviewing; however, self-administration via web and paper questionnaires represents an opportunity to lower costs, gather more timely and frequent data, and expand sample sizes to meet ever-expanding user precision requirements.

2. Embedding ART Principles in the Study Design

When considering a change from interviewer- to self-administration, several questions emerge. Key among these are the following:

- Will questions in the interviewer-administered setting translate to provide comparable data in the self-administered setting?
- Will sample members respond to the survey at an acceptably high rate? Will those who respond represent the population of interest?
- Can the survey collect high quality data (e.g., low measurement error) while leveraging the efficiencies of self-administered modes?
- What data collection protocol will yield the best overall quality given the survey goals?
- What data collection protocol will be most cost efficient?

The first question can be evaluated using a variety of pretesting methods, including cognitive interviews and online pretesting (Murphy et al. 2016; Edgar et al. 2016). To answer the remaining questions, we can design experiments in which one or more features of the data collection protocol is altered. For example, we may conduct an experiment using different incentive levels to determine which is most appropriate given the goals of the survey. Or we may randomly assign some sample members to a version of the survey with a shorter completion time and others to a longer one to determine the tradeoffs between information gained overall and from individual cases. Often, there are more issues than can be feasibly investigated in a single round of experiments. Sometimes the design options to be tested are interdependent. For example, whether incentives should be guided by response propensity models depends upon how those propensity models perform. In this case, it may be advantageous to conduct experiments iteratively, where the “best” protocol identified in one phase of the survey is carried into the subsequent phase, while the protocols that did not yield good results are excluded.

Our ability to draw conclusions from such experiments and answer questions to inform the “best” design for a survey depends on 1) the data available, 2) the specific components of quality to be monitored and 3) the interventions at our disposal to affect design features to improve quality. These three requirements can vary greatly depending on the survey mode or modes employed in the survey. For example, for a face-to-face household survey, we can measure the effectiveness of different contact strategies and interviewers, the timing and level of effort devoted to contacting respondents, the physical characteristics of the household, interviewer performance metrics and other paradata. These and other data can be monitored in real-time, analyzed to determine if an intervention is warranted and, if so, to deploy whatever intervention is indicated as quickly as possible.

Several design strategies offer the potential to help determine the best fit in terms of approaches for a survey or individual sampled units. The *tailored design* method (Luiten and Schouten 2013) advocates varying the survey protocol across population subgroups rather than using a "one size fits all" approach. This approach attempts to customize the survey design to individual preferences in order to minimize the total error for the entire sample. Another form of tailoring uses a combination of survey design features demonstrated to be effective in the literature to construct a single, optimal survey design that, when applied to the entire sample, will provide excellent results across a wide range of survey topics and populations (Dillman et al. 2014). *Responsive design* was introduced by Groves and Heeringa (2006) as an approach to adjust the data collection protocol for a single survey based on the outcomes of an initial set of cases. By continuously assessing the results of the data collection process and remaining resources, strategies can be modified for the remaining cases to be pursued (Laflamme and Wagner 2016). *Adaptive survey design* similarly proposes different approaches within the same survey, but focuses on the heterogeneity of sample cases and identifying the optimal survey protocol for each individual. For example, adaptive survey design recognizes that some sample members may be swayed to participate in a survey by incentives where others will not. Design-specific response propensities can be calculated based on paradata for each individual sample member (Schouten et al. 2017; Chun et al. 2017).

The successful ART design should adhere to these simple but key principles:

1. identify a few, critical factors that drive costs and quality (i.e., CTQs) and focus attention on these throughout the process,
2. create and monitor metrics that are strongly associated with CTQ outcomes and intervene when these metrics deviate beyond their acceptable limits and,
3. verify that the interventions were successful and that the aberrant CTQ metrics return to their acceptable limits.

A fourth overarching principle is to simplify the quality management strategy to the extent possible using informative graphical displays, parsimony in the selection of CTQs and their corresponding metrics, and a focused strategy for continual improvement of the CTQs.

These general principles are common to all three approaches – that is, A, R, and T – where the specific approach can be viewed as a variant in the way these principles are applied. For example, responsive designs may incorporate experimental phase and may use the concept of phase capacity to signal the end of a phase. These features address the principles of monitoring metrics and intervening when they meet certain prespecified criteria. Likewise, the tailored designs may attempt to adapt the data collection protocol to specific subgroups of the population. This feature can be viewed as application of

Principle 1, where metrics are defined at the subgroup level and interventions can vary by subgroup. Finally, adaptive design can contain elements of both responsive and tailored designs, but may focus more broadly on total survey error and costs, clearly in the spirit of Principles 1, 2, and 3.

ART considerations for in-person surveys are, overall, rather complex and may represent a wide array of potential CTQs for reducing error risks and costs. By comparison, the considerations for self-administered, postal surveys are relatively straightforward. For example, the typical web/paper survey involves a series of participation requests and reminders sent by mail or, if available, email according to a prespecified contact schedule with little room for deviation. The absence of interviewers and control over the timing of contact results in fewer variables to consider. However, problems can occur that may require rapid intervention when the observed results deviate substantially from expectations. In that situation, the interventions may be limited to actions such as: using additional contacts, increasing the sample size, altering the wording of the invitations, or something similar, none of which represent a substantial departure from the planned protocol. Mail invitations are typically sent in large batches and while the U.S. Postal Service (USPS) returns letters that were undeliverable, the outcome of the contact attempt is unknown unless the respondent actively participated or contacts the survey organization to refuse participation. As such, identifying, monitoring, and using data for ART designs in a web and/or mail survey environment presents a unique challenge.

Our recommended approach for surveys transitioning to self-administration is to incorporate elements of ART designs where appropriate. For instance, a design may be responsive by including multiple phases of data collection, each drawing from the successful strategies of the previous phase. The design can be adaptive in the sense that it uses paradata metrics to monitor quality during each phase of data collection to consider the appropriate treatment for each individual case. And it can be tailored in sense that it attempts to vary the survey protocol according to the (often predicted) preference of the sample member; as an example, using a paper-questionnaire-only protocol for sample members who do not have internet access.

When designing a protocol for a sequence of experiments to be conducted in rapid succession, it is vitally important to identify the goals and metrics for success from the outset. While all surveys strive for high quality in the data collected and estimates produced, the exact definition of "quality" may differ from project to project. For this reason, it is crucial for the survey stakeholders (sponsors, data collectors, data users, etc.) to discuss the definitions of quality and success from the very beginning of the survey planning stages. Once quality is defined, it is a matter of operationalizing this definition by selecting metrics that can be tracked during data collection that can serve as CTQs. These indicators can reflect quality dimensions such as successful study recruitment (response rates), the extent to which respondents represent a benchmark measurement of the population of interest (e.g., demographic characteristics that match Census estimates), success in obtaining responses at the item level from respondents, the ability to push respondents to respond via web rather than paper, and so on. Assuming a rapid development schedule, once the CTQs are identified, a system needs to be put in place to track these metrics during data collection so design decisions for the subsequent phase can be made before the current phase is complete.

The task of identifying metrics or indicators for a CTQ is seldom straightforward. As an example, the response rate is commonly employed for the CTQ to minimize nonresponse bias; yet, other metrics maybe better indicators of nonresponse bias. Often the best solution is to employ several metrics that may reflect different dimensions of the CTQ but that can add complexity to the monitoring task. Thus, it is important to strike a balance between parsimony and completeness. In the RECS pilot studies, we monitored response rates as well as a measure of representativity known as Cramér's V. In addition, it may be futile to define a CTQ for which there is no opportunity or plan to intervene on behalf of the CTQ. As an example, in the RECS pilot studies, obtaining accurate reports of household appliances was certainly an objective of questionnaire design; we refined these questions during pretesting and checked the data for anomalies at several points during data collection, but no steps were taken to monitor this indicator on a daily basis.

In the following section, we present a case study to illustrate a rapid sequence of experiments for the RECS that incorporate various ART design principles. We describe our process for identifying CTQs and monitoring them during data collection using a visualization system designed for such scenarios.

3. The Need for Rapid Decision-Making and Role of Visualization for the RECS Pilots

The RECS originated in 1978 and has been conducted periodically by the EIA since then. The RECS program is responsible for collecting and disseminating timely, detailed information about how energy is being used within the residential sector of the economy. This includes data on the fuels used in homes, equipment and appliance stocks, household behaviors, and disaggregated consumption and expenditures. In this article, we discuss the RECS that was conducted in 2015. Prior to this, RECS was conducted by face-to-face interviewing in 2009. As the opportunities and challenges associated with survey research have changed over the years, the planning for each RECS has required reflection on how best to meet the goals of the survey and needs of the data users while maintaining comparability with past rounds, and adhering to schedule and budget constraints.

As previously noted, the RECS has been conducted by field interviewers using computer-assisted personal interviewing (CAPI). The costs of this data collection mode are relatively high, averaging nearly USD 400 per completed interview for the 2009 study, which limits other important quality initiatives, such as more frequent data collection, larger sample sizes, and precise estimates for more geographic areas. EIA commissioned an expert panel study of the National Research Council of the National Academy of Sciences (NAS) to examine its energy demand surveys, identify gaps in substantive coverage, and make recommendations for EIA's priorities for data collection (Eddy and Marton 2012). One suggestion from the panel was to explore alternative data collection approaches for the RECS, specifically incorporating a self-administered web mode. EIA followed the NAS recommendation with the goal of assessing whether self-administered modes could result in the collection of high-quality data in an environment where field data collection was continuing to face increasing challenges.

RECS had always achieved response rates at or above 80 percent. It was apparent that such high response rates would not be feasible in today's environment, especially using a mail-delivered questionnaire protocol. Thus, EIA faced many questions about the trade-off between cost and quality with this radical shift to a web/paper design. The schedule for conducting the next RECS meant that EIA would need to test several different promising protocols in a very short period of time before selecting the one that would serve as the most appropriate production system for the future. In light of these challenges, EIA determined that, using the ideas of responsive design, a multi-phased pilot study design would best meet their needs considering costs, timing and goals. With a phased approach, a well-selected set of design features could be tested at each phase that took advantage of the lessons learned and the results gleaned from the prior phase's experiments. Then the final phase of testing could incorporate the best features of the prior phases.

The following sections describe each RECS pilot phase's timing, design, and results. We also describe how early results from each phase were used to inform subsequent phases, as well as how the official 2015 RECS CAPI study was ultimately impacted by the pilot tests.

3.1. Phase 1: The Cities Pilot

The first RECS phase, referred to as the Cities Pilot Test, collected responses from an address-based sample of households in five U.S. cities. Planning and design for the Cities Pilot Test began in December 2014 and data collection began in March 2015, continuing into July 2015. Planning for this survey involved extensive cognitive interviewing and pretesting to determine how best to shorten the 40-minute traditional face-to-face questionnaire to a 20–30 minute self-administered instrument (Murphy et al. 2016). Further, extensive analysis was conducted on the energy characteristics of U.S. cities in order to identify five cities that together could sufficiently represent the diverse and challenging issues to be confronted in the redesign of the national RECS.

In addition to assessing the viability of a self-administered RECS, the Cities Pilot included two experiments to evaluate options for key design components. These components were (1) questionnaire length and (2) initial mode assignment. We found evidence that the 30-minute self-administered RECS achieved a similar response rate to the 20-minute version and deemed it feasible and efficient for both web and paper modes. The Cities Pilot Test achieved a higher-than-expected response rate overall (38%) within budget and demonstrated that data collection can be accomplished within a 14-week field period. We also found that tailoring the initial mode assignment (either by web only or by paper) based upon a model predicting the propensity to respond by each mode was not effective primarily because our working hypothesis did not hold. We hypothesized that households that do not have broadband Internet access would prefer the paper mode. Thus, we developed a model for the probability a household has Internet access as described in Zimmer et al. (2016). But while the propensity model was reasonably accurate for predicting Internet access, Internet access was apparently not a good indicator of mode preference and thus response rates for the model-guided protocol were not significantly higher than the control group which used a static web-first mode assignment (Zimmer et al. 2016). Given these results, mode tailoring based on internet access propensity was not used in the subsequent phases.

As shown in Figure 1, the Cities Pilot Test data collection phase overlapped significantly with planning for the next phase, referred to as the National Pilot Test. Daily tracking of key Cities Pilot quality metrics was instrumental in determining design and experiment options for the subsequent phase. Monthly, or even weekly, status reports would have been insufficient if the project were to stay on schedule.

3.2. Phase 2: The National Pilot Test

The RECS National Pilot Test was planned to run more or less concurrently with the official 2015 RECS CAPI study and, thus, planning and decision making for the National Pilot needed to take place while the Cities Pilot data were still being collected. Planning and design for the National Pilot Test began in April 2015; data collection ran from September 2015 to January 2016. The RECS National Pilot Test collected responses from a national address-based sample of households and expanded on lessons learned from the RECS Cities Pilot, carrying over the 30-minute questionnaire length and materials developed for that previous pilot.

Like the Cities Pilot Test, the National Pilot Test phase included experiments to continue to explore the most successful protocol for web and paper administration of the RECS according to the criteria outlined in Section 4 of this article. Because the Cities Pilot Test showed evidence of the superiority of the web mode for data collection in terms of cost and data quality, we aimed for a design that would effectively “push” respondents to the web (Dillman 2016). Since respondents in the Cities Pilot generally preferred to respond via paper rather than web, a key question for the National Pilot Test was whether participants could be incentivized to respond by web rather than paper.

The pilot evaluated eight treatment combinations of equal sample size formed by crossing two factors: respondent incentives (Factor A) with two levels and mode protocols (Factor B) with four levels forming a two-by-four factorial design. The two factors and their levels are defined as follows:

- **A1.** A USD 5 prepaid incentive included in the first questionnaire mailing; USD 10 was promised for response under the response protocol specified by Factor B.
- **A2.** A USD 5 prepaid incentive was included in the first questionnaire mailing; USD 20 was promised for response under the response protocol specified by Factor B.
- **B1.** Web Only Protocol−only the web response option was offered for all survey response invitations.

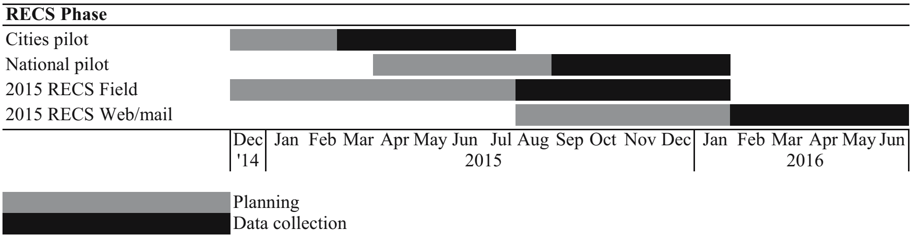


Fig. 1. RECS pilots timeline.

- **B2.** Web/Paper Protocol – the web response option was offered in the first invitation and first nonresponse invitation; both web and paper were offered in all subsequent invitations.
- **B3.** Choice Protocol – response by either paper or web questionnaire was requested by each survey response invitation.
- **B4.** Choice+ Protocol – response by either paper or web questionnaire was requested by each survey response invitation. However, a USD 10 promised bonus incentive was provided in addition to the incentives specified by Factor A if the respondent chose to respond by the web option rather than by paper.

The National Pilot Test found that the most successful protocol included a USD 5 unconditional cash pre-incentive plus a USD 10 cash promised incentive for participation. Those electing to respond via web rather than paper were offered an additional USD 10 for completing. This protocol, termed "Choice Plus" (Choice+) is discussed in detail in Biemer et al. (2017b).

Following the main data collection period of the National Pilot, all non-respondents except refusals received an extended nonresponse followup (xNRFU). A single UPS high-priority mailing was sent to these addresses containing the offer letter, an abbreviated, one-page questionnaire and a postage-paid return envelope. A random half of the xNRFU sample was offered an additional (that is, in addition to the incentives they would have received under Factors A and B) USD 10 if they completed the abbreviated questionnaire and returned it in the stamped envelope.

We calculated response rates using American Association for Public Opinion Research formula RR3 (AAPOR 2015). The final overall response rate for the main phase of the National Pilot was 40.4 percent. The overall rate rose to 54.9 percent after the completion of the nonresponse follow-up phase.

3.3. Phase 3: 2015 RECS CAPI Remediation

While the National Pilot was being conducted, the 2015 RECS CAPI was also in the field with interviewers making face-to-face visits to selected households. The National Pilot Test ran concurrently with the 2015 RECS CAPI so we could compare the results of data collection and estimates obtained by survey mode. In January 2016, it became apparent that the 2015 RECS would not achieve its goals in terms of cost and response using interviewer administration. In contrast, the National Pilot had concluded and demonstrated that self-administered modes could be used to achieve good data quality, acceptable response rates, and costs several times lower per case than CAPI. Given the success of the National Pilot Test to that point, and in particular the Choice+ protocol, EIA made the decision to transition most unresolved or unreleased sample cases in the 2015 RECS CAPI to self-administration by web and paper using the Choice+ protocol beginning in February 2015. The 2015 RECS CAPI remediation phase continued through June 2016.

4. Identifying and Visualizing CTQ Metrics

Historically, RECS had a fairly stable design with a predictable range of outcomes. The response rate for the CAPI studies, for example, was consistently 80 percent or higher.

To monitor field progress during data collection, RECS project staff relied on summary field and cost tables which were compiled at regular intervals. These static tables, which included weekly and cumulative labor and travel costs for the entire sample and by geography, were sufficient. The use of more detailed metrics was limited to field supervisors as a means to appropriately assign interviewer work. Given its quadrennial cycle, there was also ample time to analyze field results following data collection and plan for any protocol changes for the next round.

The objectives of the RECS Pilot Test required rapid decision making in order to plan the subsequent round, and it was necessary to track data collection metrics from the outset and at more frequent intervals. As the process evolved, it was essential to assess the current state of progress for CTQs in a way that did not require a significant time investment from staff. Making design decisions on the schedule presented in Figure 1 required real-time (daily) monitoring of the performance across multiple quality indicators.

To develop a quality monitoring system, the first step was to agree upon the definition of quality for the pilot tests. A number of design features were predetermined such as the data collection modes (web/paper), overall sample sizes, the use of an address-based sampling frame, number of mailings and incentivized response. Given these fixed assumptions, the definition of quality and what constituted a successful outcome from the survey sponsor's perspective drove the remaining design decisions. Thus, quality was essentially defined as a balance across the following four CTQs:

- participation rates (higher is better, all else being equal),
- web response rate (higher response by web compared to response by paper is better),
- respondent sample representativeness relative to external benchmarks (higher concordance with benchmark is better), and
- relative costs per completed case (lower is better).

Another important quality goal for any survey transitioning from face to face to web/mail mixed mode data collection should be the evaluation and control of mode effects which are essentially the methodological differences in the estimates due to the change in mode. The RECS is certainly not immune to mode effects; in fact, significant mode effects were expected for some characteristics, most notably the ascertainment of housing unit square footage. In the face to face mode, interviewers can explain the "official" concept of housing unit square footage and even assist the respondent in estimating it. In self-administered modes, respondents are only aided by the instructions embedded in the instrument which can be quite technical. Unless they happen to know the square footage of their home, respondents often err in its estimation. Unfortunately, measurement errors are quite difficult to monitor and control in real-time during web/mail data collection. For the RECS, only post-survey evaluations of measurement error were conducted. In particular, a post-survey analysis of RECS square footage data can be found in Amaya et al. (2017).

While we did not set a quantitative value for each of these CTQs to identify what worked "best," we monitored the results as the pilot tests were conducted and frequently discussed the trends in the process of selecting methods appropriate for the subsequent pilot. Biemer et al. (2017b) provides a discussion of the specific rationale for the chosen survey experiments for the RECS National Pilot design to identify the "best" design given the definition of quality above.

The CTQs we identified for real-time monitoring can be classified into the following categories:

1. *Participation.* We sought high rates of participation across key domains defined by housing unit/household characteristics. We also sought high rates of participation in the early stages of data collection to minimize the cost of multiple follow-up mailings. We also tracked submission rate metrics for each of the experimental conditions. Here, submission rate refers to the count of cases submitted via web or paper form divided by the total number of sampled cases. The rate served as a simple proxy for response rate, which was not calculated until the end of data collection due to the timing of defining criteria for the estimation of eligibility among cases of unknown eligibility (e.g., cases with no evidence of receipt of contact with a respondent, USPS and UPS undeliverables). Cumulative daily submission rates were monitored overall and by: experimental treatment, mode protocol, promised incentive amount, geographic region, and urbanicity.
2. *Rate of response via web (rather than paper).* We sought to minimize cost by encouraging response via web survey rather than paper, since paper included extra costs for printing, return postage, receipt, data entry, and data review. We also sought to minimize measurement error from item nonresponse, out-of-range responses, and errors in following skip patterns by encouraging web vs. paper response. We monitored the rate of web submission overall and by the same factors noted for participation.
3. *Respondent representativeness compared with the sample and an external benchmark.* We sought a balanced unweighted distribution of respondents relative to benchmark data sources. To assess representativeness, we compared RECS Pilot responding housing unit/household distributions to the corresponding distributions for the 2014 U.S. American Community Survey (ACS) 1-year estimates (U.S. Census Bureau 2015). The variables compared included: type of housing unit, main heating fuel used, household income, and age of respondent/householder. Additionally, we compared respondents to sampling frame distributions using a variable available for both: housing unit building type (single family vs. multi-unit). We also considered comparing RECS Pilot respondents with 2015 RECS CAPI respondents or 2009 weighted estimates on energy-specific metrics, such as water heating fuel and number of refrigerators, as a means to track the bridge between old and new survey methods. These metrics were not part of the CTQ tracking system, but rather were evaluated at the conclusion of the data collection.
4. *Relative cost per case.* We calculated the costs associated with printing materials, mailings, receipt of completed questionnaires, and incentive payments for each protocol. Depending on whether and when each sample member responded, the cost per case varied. By tracking costs at the case level, we could determine the overall costs at the protocol summary level over the course of data collection. We measured costs relative to the simplest and expected least costly protocol, Web Only. We also considered the costs associated with data editing needs for each protocol. Data editing involved necessary recodes to reported values, such as the assignment of values to open-ended responses or edits to ensure consistency of responses. The need

for data editing did not necessarily suggest lower quality data, but did require staff resources, and so we discuss it as a cost metric.

With our CTQs for the pilot tests identified, the next challenge was selecting the best system for monitoring progress data collection. In selecting an approach for CTQ visualization, we identified several system requirements specific to our purposes:

- The system should be simple enough to be quickly accessed and understood by a wide range of project staff.
- Charts should limit the number of data series plotted (no more than five lines on a single chart at once; as needed, "small multiples" of charts by specific dimensions should be employed (Cleveland 1993; Tufte 2001).
- The chart axes, legends, and labels should be large enough to read easily and include descriptive labels to minimize the effort for a user to gain information.
- Charts should use a consistent format and consistently use patterns, colors and symbols so users do not have to re-orient when examining multiple charts.
- The charts should be fully interpretable when printed in black and white.
- All charts should be accompanied by full data tables so users can reference exact values when needed.
- The charts should not require purchase or licensing of special software not already available to users.
- The system had to be cost efficient and not require extensive use of IT resources.

We determined that the goals stated above could best be met using software already at our disposal – SAS and Microsoft Excel. The creation process began with a nightly export of data from the survey control system into a single SAS database containing sample, frame, case history, web response data, paper response data, and auxiliary data. SAS version 9.4 was used to compute CTQs in counts per day, cumulative counts per day, and cumulative rates per day for each metric. The same automated SAS program then prepared data tables in the format necessary for export to an Excel 2013 workbook with preformatted chart shells.

The process was automated to run on a daily basis for sharing with the project team on a shared secure web portal. By containing all the output in a single Excel workbook, we had available a self-contained, convenient, and widely familiar format for sharing pilot progress data across the project team and with EIA managers, as needed. And although the charts were not necessarily "interactive," it should be noted that Excel does include the default option to hover the cursor over a data point to view its exact x and y values.

For chart designs we looked to the literature and our own experience and intuition about the simplest and most effective displays. For example, Tufte (2001) advocates for maximizing the "data-ink ratio" or "proportion of a graphic's ink devoted to the non-redundant display of data-information" in data visualization. As a result, we sought to eliminate any elements that were not helpful for quick interpretation of the data such as excessive gridlines, non-meaningful uses of color, redundant labels, and so on. We consciously adhered to other evidence and advice from documented best practices of data visualization regarding the choice of chart types. For instance, for our cumulative charts, we used line charts connecting individual numeric data points, which have been advocated as a “simple, straightforward way to visualize a sequence of values. Their primary use is to display trends over a period of time” (Hardin et al. 2012).

Figure 2 presents a “snapshot” of the RECS National Pilot monitoring display referred to as an Adaptive Total Design (ATD) monitoring chart (see, for example, Biemer 2010), formatted using a dashboard-type design. The legend at the top of the figure describes the markers used on the x-axis in this and subsequent figures to identify key dates in the data collection protocol. Having this information consistently displayed daily with various metrics in close proximity allowed our team to quickly ascertain and keep apprised of the various CTQs and their performance throughout data collection. The charts included a mixture of line (for time-dependent, cumulative rates), bar charts, and maps. Though it is not a standard option for Excel, maps are not difficult to add to a workbook using Visual Basic for Applications (VBA) (Camoes 2008).

To illustrate an individual chart from the dashboard, in Figure 3 we present cumulative submission rates by protocol during the National Pilot Test data collection period. At a glance, it is obvious that submission rates rose most rapidly at the beginning of data

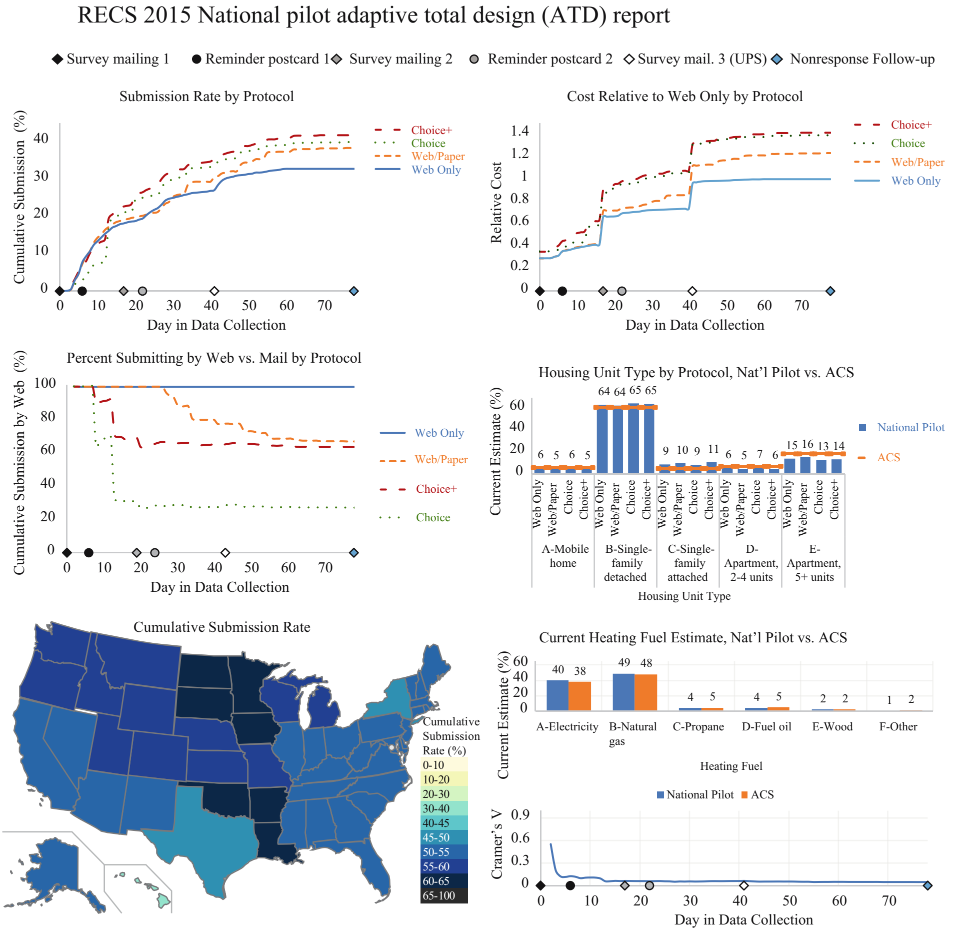


Fig. 2. National pilot ATD monitoring charts (partial view).

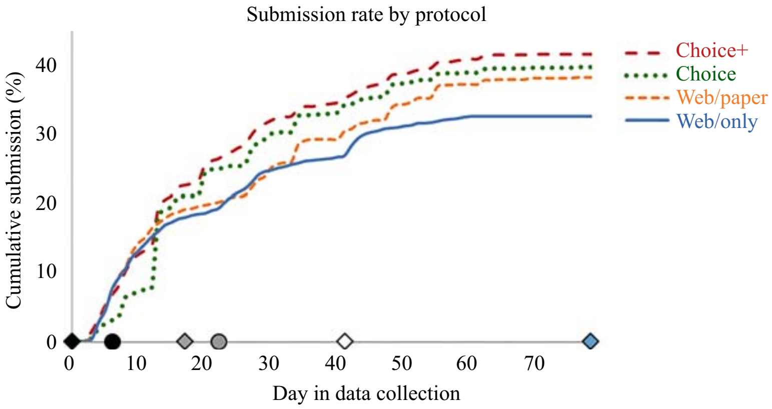


Fig. 3. National pilot submission rates by protocol.

collection in response to several mailings to sampled households. A second spike occurred just after the nonresponse followup mailing (diamond at day 78) among all experimental protocol groups, but the increase was most dramatic for the Web Only group that had not previously been offered a paper option for response. Figure 3 exemplifies the format of our charts. Reviewing this chart regularly during data collection was critical for identifying the best protocol for use in the second phase of the 2015 RECS.

As evident in Figure 3, all protocols performed similarly in terms of submission rate prior to the nonresponse follow-up period, with the possible exception of Web Only. While this version of the daily monitoring chart does not, reflect the statistical uncertainty (e.g., standard errors) around estimates, we recognize the importance of including this information when comparing protocols. We designed the experiment such that both groups had a robust and equal sample size (2,412 in each protocol). This means that a practical difference of a few percentage points would be statistically significantly different as well. We calculated and check for statistically significant differences at certain "check points" during data collection. In Figure 4, we present a version of the chart monitoring Choice+

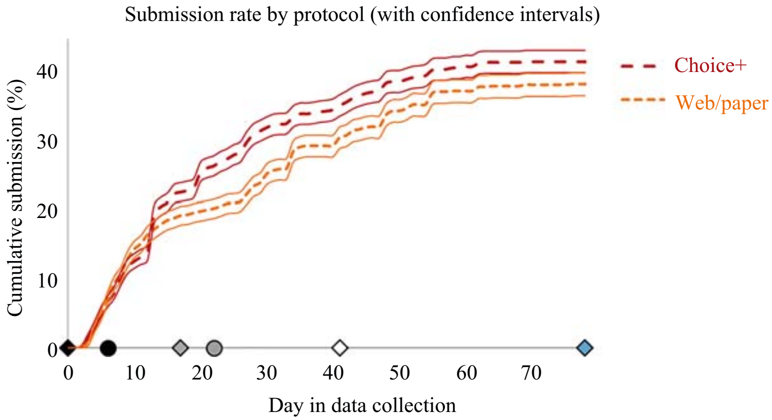


Fig. 4. National pilot choice+ and web/paper submission rates with 90 percent confidence intervals.

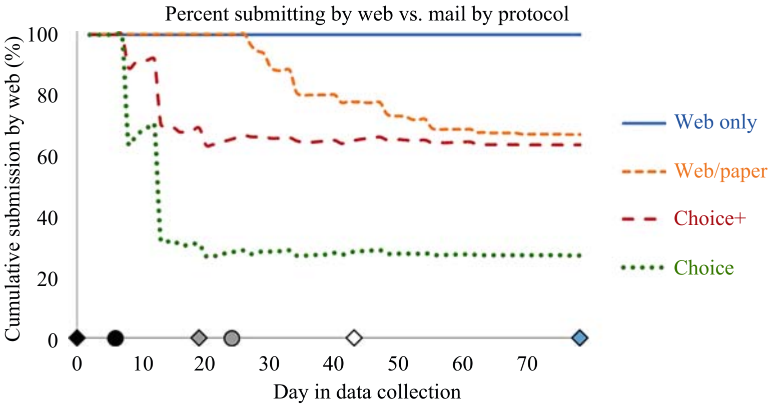


Fig. 5. RECS National pilot cumulative proportion of interviews completed via web.

and Web/Paper submission rate with the inclusion of the 90 percent confidence intervals for each (to test for whether the Choice+ submission rate was significantly higher than Web/Paper). To render these lines, we computed for each protocol and day in the data collection period using the formula for confidence intervals for a one sample dichotomous outcome.

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{1}$$

The upper and lower confidence interval values for each protocol were added to the chart. This made it possible to see where those intervals overlapped for the two protocols. Figure 4 shows the difference between the protocols' submission rates becoming significant around day 20 in the data collection period. They remained significantly different through the end of the main data collection phase, though the rates were only different by a few percentage points.

Submission rates represent one dimension of quality, but given that a key concept of ART design is monitoring multiple indicators of quality and cost, we cannot rely on overall submission rates alone. For instance, it was noted that a significant cost and quality driver was the proportion of interviews that would be completed via web surveys as opposed to paper. As shown in our next chart example (Figure 5), both Web/Paper and Choice+ had a majority of interviews completed via web during Phase 1 of the National Pilot. The Choice group resulted in only about a third of cases being completed by web. The Web Only group, by definition, had 100 percent of cases completed by web during Phase 1. Taken together with submission rate, these two metrics begin to paint a more complete picture of the quality and cost tradeoffs of the different treatments. It was only because we were tracking these trends closely that we could make the rapid decision to implement a self-administered protocol (Choice+) for the 2015 RECS.

Another CTQ metric that was closely monitored reflected the ability of each protocol to elicit responses from key respondent domains. One such characteristic of interest was the age of the householder – a characteristic we expected to have some correlation with mode preference as web access and use tends to be higher among younger individuals. As a benchmark, we used estimates from the 2014 ACS for householder age in our sampled areas. In Figure 6 we present our chart for monitoring the distribution of respondent

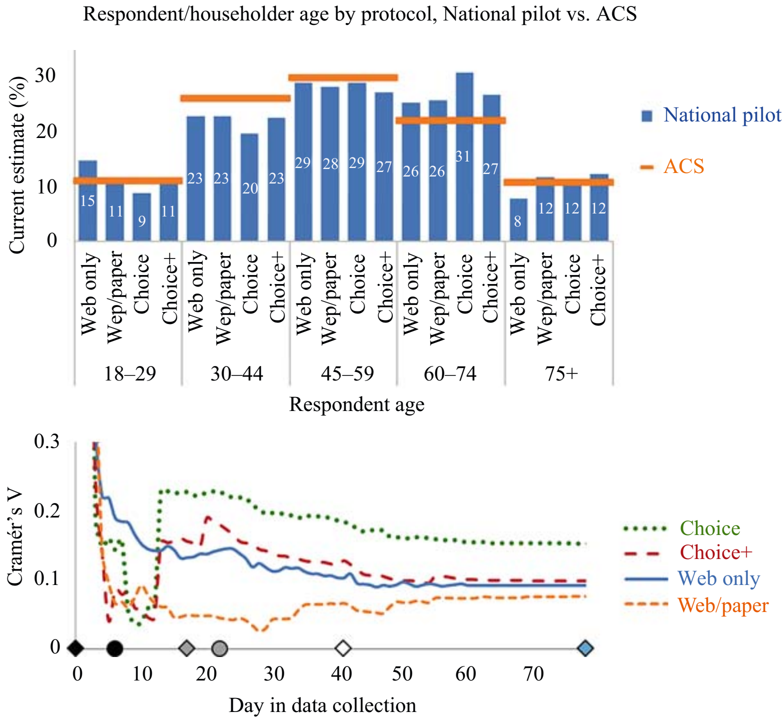


Fig. 6. RECS National pilot age of respondent vs. ACS Age of householder.

(assumed to be the householder) age during data collection. The bar chart at the top shows that no protocol resulted in a distribution matching the ACS exactly, but the differences were not extreme.

To monitor householder age and representativity trends over the course of data collection, we opted to compute a single index to communicate the ability of the protocol to realize a sample matching the ACS. We used the Cramér's V measure, which can be used to determine the degree of association between nominal variables (Cramér 1946). The value ranges from 0 (no relationship) to 1 (perfect relationship) with values under 0.2 indicating a very weak relationship. In this case, the relationship is between the age distribution and the surveys (National Pilot Test and ACS) so a weak relationship suggests little difference between the surveys (i.e., lower is better).

Several other measures of representativeness were considered including the dissimilarity index (see, for example, https://en.wikipedia.org/wiki/Index_of_dissimilarity) and various sample balance indicators such as the R-indicator (Schouten et al. 2009). The former provides essentially the same information as Cramér's V and its advantages over V is only a matter of personal preference and a similar discussion of the post-hoc use of the dissimiliary index can be found in Biemer et al. (2017b). However, the latter measures are inappropriate for comparing a sample variable distribution to an external benchmark distribution which was the objective of our representativeness criterion.

It should be noted that high values of V do not necessarily mean the protocol will result in biased estimates. The respondent sample was ultimately adjusted for nonresponse and coverage error, correcting some of the non-representativity. In addition, perfectly representative samples do not always generate V = 0. The 2014 ACS was conducted nearly two years before the National Pilot and any changes in the target population during that time could cause an increase in V. Also, both the National Pilot and ACS are subject to sampling error and this has not been taken into account in this analysis. Finally, minor differences in the wording of questions, eligibility criteria for housing units, and target populations for the two surveys could impact V. Regardless, we found Cramér's V useful for highlighting practical differences in representativity among the protocols.

The bottom half of Figure 6 shows the cumulative Cramér's V values for age distribution by protocol. Once the survey was several days into data collection, there was little difference between the National Pilot Test and ACS age distributions and the difference became smaller over time as more interviews were completed. The largest difference between the National Pilot Test and ACS by protocol was seen in the Choice group. A review of the top half of Figure 6 suggests that National Pilot Test Choice respondents skewed towards the older age groups, suggesting that the Choice protocol, relative to the other protocols, was on the whole a more attractive option for older respondents. A review of mode choice suggests that older respondents were much more likely than younger respondents to select the paper mode and the Choice protocol did little to dissuade respondents from choosing paper over web. As shown earlier in Figure 5, however, Choice+ offered an additional incentive for web response and was much more effective at attaining a high proportion of completed via web compared to paper during the main data collection phase.

Regarding costs, we compared the average cost per case across the data collection period by protocol. We began with the understanding that any self-administered protocol would be several times less expensive than CAPI, but were interested to compare costs between self-administered protocols to inform future designs. While cost was not the primary concern in comparing self-administered protocols, it was an important dimension. In Figure 7, we present the cost per case of each protocol, relative to Web Only (with a value of 1 at the end of the data collection period). For each protocol, there was a large increase in costs with each subsequent mailout (sent only to nonrespondents). Between mailouts, costs

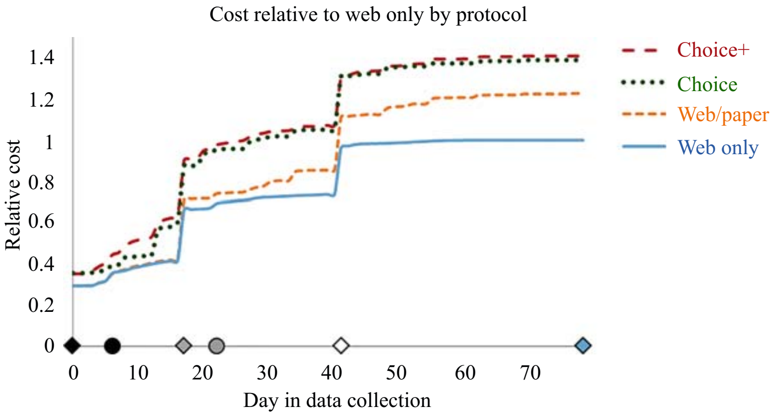


Fig. 7. RECS National pilot relative cost per case by protocol.

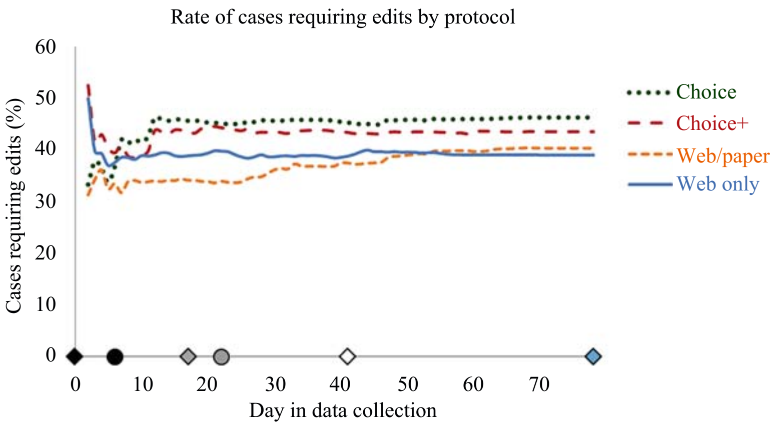


Fig. 8. RECS National pilot rate of cases requiring edits by protocol.

were incurred with the receipt and entry of paper questionnaires and payment of incentives. In the end, we found Web/Paper to be about 20 percent more expensive than Web Only and Choice and Choice+ to be about 40 percent more expensive.

Much of the content of the RECS questionnaire is technical, therefore considerable staff time can be devoted to reviewing inconsistent or improbable responses, such as lack of heating equipment in cold climates or extremely large housing unit measurements. Editing, therefore, is considered as much of a cost metric as it is a quality one. Figure 8 presents the rate of completed interviews requiring data edits by protocol. All responses, regardless of mode, were subjected to the same edit specifications. However, the rate for Web Only was lowest since the web allows for greater restriction of response options and programmatic skips through the questionnaire. The rate of data edits required for the other protocols, which included paper responses, were higher. Web/Paper initially had the lowest edit rate, but as the later period allowed for paper questionnaires to be submitted, this rate increased.

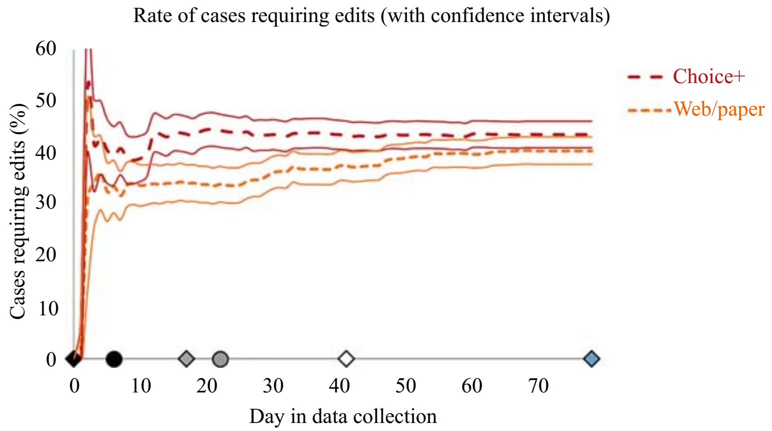


Fig. 9. National pilot choice+ and web/paper editing rates with 90 percent confidence intervals.

In Figure 9, we add one-sided 90 percent confidence intervals to compare the editing rate for cases in the Choice+ and Web/Paper protocols, testing to determine whether the Web/Paper editing effort was statistically significantly less expensive than Choice+ over the course of data collection. The figure suggests that very early in the data collection period, the difference in rates was not statistically significant, but that after the second week, enough cases had been completed to determine a difference. However, later in the data collection period, after the Web/Paper protocol introduced the paper option, the rate of cases requiring editing rose for that protocol so that the rate was no longer statistically significantly lower than that for Choice+.

Taken together these metrics illustrated in Figures 3 through 9 begin to paint a more complete picture of the quality and cost tradeoffs of the different treatments. Reviewing these charts regularly during data collection was an important step in identifying the best protocol for use in the 2015 RECS CAPI remediation. It was only because we were tracking these trends closely that we could make the rapid decision to implement a self-administered protocol (Choice+) for the 2015 RECS CAPI. Choice+ demonstrated the ability to achieve the highest level of response, a majority of cases responding by web vs. paper, good comparability with external benchmarks, and costs that were reasonable for the needs of RECS.

As a final example, we include in Figure 10 a monitoring chart helpful for decision making during the RECS 2015 CAPI Remediation Phase. By monitoring data collection progress regularly, we could follow the submission rate trend in all phases of the Pilots to determine that the self-administered protocol was achieving a similar or higher rate in a shorter amount of time. This chart helped identify and communicate the impetus for switching to web/paper for the remediation.

Once the charts were produced, it was important to get them in the users' hands to facilitate discussion and planning. Our nightly process published the charts in a single file on the secure project web site where all users could access and download the file. We referred to the charts in day-to-day planning and included a copy with the materials for each of our weekly planning meetings. We found this approach minimized the burden on individual users while maximizing the reference to and use of the charts for decision making.

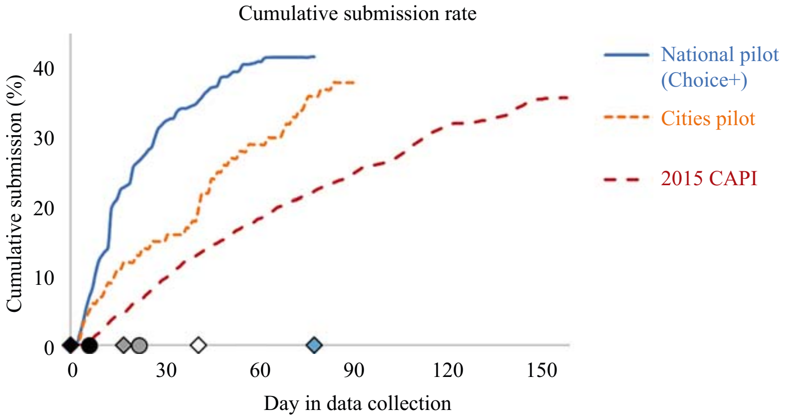


Fig. 10. RECS CAPI Remediation submission rates by phase.

5. Discussion

The ART approaches described in this paper proved to be quite powerful for the RECS and were essential for guiding the experimentation and field work during the piloting of self-administered modes, then the transformation from face-to-face to web/paper administration. There are several key lessons learned, however, and it may take multiple survey cycles to develop the right mix of CTQ metrics for RECS. With the RECS Pilot, we focused primarily on data quality because the approximate cost savings of moving from CAPI to web/paper were easy to predict. In future rounds, cost will be a much more critical metric to track, as EIA continues to explore the optimal mix of web/paper or web/paper/CAPI modes. In particular, project staff might trade lower costs for greater data quality in the design of experiments to determine the proper mix and sequencing of modes, whether to use non-English survey instruments, when to implement stopping rules, and the scope of nonresponse follow-up. In addition, suggestions for CTQs should be solicited from staff involved in downstream processes such as data editing, weighting, imputation, and energy modeling Finally, as suggested in the previous section, the scope of the CTQ metrics should include more energy-specific comparisons using the prior RECS estimates as benchmarks. These comparisons would balance the metrics tracking demographic representativity with the energy characteristics representativity of RECS respondents.

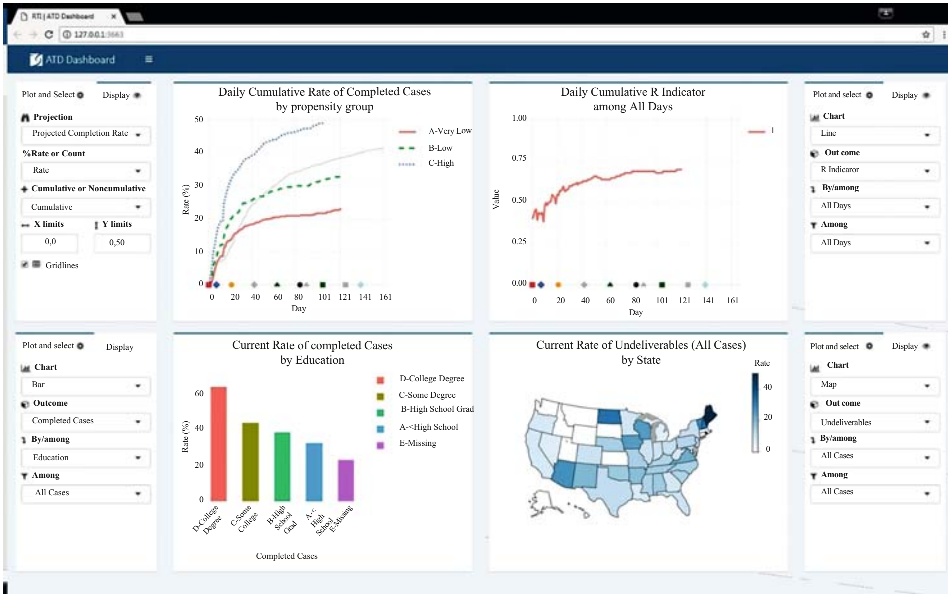
As previously noted, the successful implementation of ART designs requires monitoring critical metrics in real-time and extrapolating current trends to accurately predict future outcomes. These predictions become the basis for designing effective and timely interventions that minimize survey costs, mitigate the highest error risks and avoid major schedule delays. Visual displays of trends in the performance data supplemented by statistical tests of significance allows survey managers to detect the indications of anomalies that require action early on and in real-time when such actions are the most effective.

Our basic approach is generalizable to virtually any survey facing similar transformative decisions based upon a sequence of experiments that must be conducted in rapid succession with little or no time for analytic pauses between data collection phases. Notwithstanding the success of our current approach, there are several important ways visual ATD can be improved by the addition of features, options and tools that would enhance its utility while improving its functionality.

1. **Interactivity.** We are currently embedding interactive functionality in the ATD system (Murphy et al. 2017; Duprey et al. 2017). In particular, interactive visualizations are very useful to detect data anomalies and/or interactions among error sources and to search for their probable causes. Users are presented with an array of display options and mechanisms for categorizing, subsetting, and aggregating data, as well as overlaying projections, survey outcomes from prior rounds, or model-derived predictions. Given that data inputs may be derived from disparate systems and may exist at multiple units of analysis (e.g., sample-member level, interviewer-level, day level), a data taxonomy embedded in the display and selection menus that restrict combinatorial structures to only logical instantiations is also being implemented. Thus, CTQ indicators can be prominently displayed while extraneous information is minimized, using best practices of visual design (see, for

   example, Cleveland 1993). Figure 11 provides a snapshot of the interactive system under development.

2. **More Predictive Metrics.** Ideally, a metric for a CTQ is one that can accurately indicate when the CTQ falls below a quality level where some remedial intervention is required to achieve a desired or specified output quality level. Although good metrics exist for some CTQs such as response rates, standard errors and sample balance, this is not true for other sources such as mode effects and other measurement errors data validity/reliability and questionnaire design flaws. For field studies, we are adding visualization metrics based upon computer assisted recorded interviewing (CARI) to detect interviewer errors due to poor interviewing performance, fabrication, violations of protocols and the like. Similarly, CARI metrics can be devised to detect respondent comprehension issues or questionnaire flaws that cause confusion during the interview. Going beyond CARI, it may be possible to embed a limited number of replicate measurements in the instrument to detect response reliability and validity issues. Consistency checks can also be used to detect some types of measurement errors. For example, a model derived estimate of square footage based upon number of rooms, floors, inclusion of attics, basements, etc. could be used to identify gross errors in the estimation of housing unit square footage. These metrics would supplement and enhance the CARI metrics and other traditional metrics based upon response patterns (such as straight-lining) and response latency.
3. **Interpreting Variation**. An important issue in the interpretation of visual information is separating variation that is inherent in the data collection process (referred to as "common cause") from variation that is due to anomalous stimuli (referred to as "special cause"). It is important to distinguish between common and



Projection: Projected Completion Rate
%Rate or Count: Rate
Cumulative or Noncumulative: Cumulative
X limits: 0,0
Y limits: 0,50
☑ Gridlines

Chart: Line
Out come: R Indicaror
By/among: All Days
Among: All Days

Chart: Bar
Outcome: Completed Cases
By/among: Education
Among: All Cases

Chart: Map
Out come: Undeliverables
By/among: All Cases
Among: All Cases

Fig. 11. Interactive monitoring dashboard example.

special cause variation because their mitigation strategies are distinctly different. Special cause variation can be addressed by targeted interventions while common cause variation is mitigated by redesigning the process. Methods for interpreting variation are well-known in the quality control literature (see, for example, Breyfogle 2003). Morganstein and Marker (1997) and Biemer (2010) describe how these methods can be applied to survey processes. Adding these features to the ATD system is a priority because of the risks to survey costs and data quality of misinterpreting and inappropriately mitigating temporal and spatial variation.

4. **Automatic Detection of Anomalies**. The age of "big data" has brought about an explosion in the volume, velocity and variety of data available for anomaly detection. We have already seen an explosion of paradata and their associated metrics for detecting a wide variety of cost, quality and data timeliness anomalies. These will increase exponentially as the search for anomalies extends to interviewers and respondents at varying levels of geography, for a variety of questionnaire items, cross-classified by interviewer, respondent and geographic characteristics. The search for anomalies in the data is made even more complex by the need to identify special versus common cause variation. Fortunately, artificial intelligence provides a solution for competently managing these data at lightning speeds to detect data problem early in real time. We believe the automatic detection of data anomalies is a high priority because managing these data complexities, detecting actionable patterns in the data and prioritizing apparent anomalies according to their error costs and error risks all in real-time and with high accuracy will not be possible without it.
5. **Usability Research**. We have observed that the visual ATD system worked well for the goals of the RECS project. However, we have yet to carefully evaluate the process by which users interpret the charts and whether those interpretations are accurate. It is important to avoid the situation where users rely on fast, instinctive and emotional thinking to draw conclusions from the graphics (Kahneman's (2011) "System 1") and support the slower, more deliberative, and more logical thought process of users ("System 2"). By evaluating users' interactions with the visualizations and assessing their usability relative to alternative visualizations (Hornbaek and Frokjaer 2003), we can improve the design, resulting in even more effective interpretation and decision making.

6. References

AAPOR (The American Association for Public Opinion Research). 2015. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, 8th edition. Oakbrook Terrace, IL: AAPOR.

Amaya, A., P. Biemer, and D. Kinyon. 2017. "Total Error in a Big Data World with Applications to the Residential Energy Consumption Survey." Presented at the American Association for Public Opinion Research Annual Conference, New Orleans, LA.

Biemer, P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5): 817–848. Doi: https://doi.org/10.1093/poq/nfq058.

Biemer, P., K.H. Harris, B. Burke, K. Considine, C. Halpern, and C. Suchindran. 2017a. "Transitioning an In-Person Longitudinal Survey to a Mixed-Mode, Two-Phase Survey Design: Preliminary Results." Presented at the Annual Conference of the American Association for Public Opinion Research. New Orleans, LA.

Biemer, P., J. Murphy, S. Zimmer, C. Berry, G. Deng, and K. Lewis. 2017b. "Using Bonus Monetary Incentives to Encourage Web Response in Mixed-Mode Household Surveys." *Journal of Survey Statistics and Methodology*. Doi: https://doi.org/10.1093/jssam/smx015.

Breyfogle, F. 2003. *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. Hoboken, NJ: John Wiley & Sons.

Camoes, J. 2008. *How to Create a Thematic Map in Excel*. Available at: http://www.excelcharts.com/blog/how-to-create-thematic-map-excel/ (accessed November 26, 2017).

Chun, A.Y., B. Schouten, and J. Wagner. 2017. "JOS Special Issue on Responsive and Adaptive Survey Design: Looking Back to See Forward – Editorial." *Journal of Official Statistics* 33(3): 571–577. Doi: http://dx.doi.org/10.1515/JOS-2017-0027.

Cleveland, W. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Dillman, D., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, and Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 4th Edition. Hoboken, NJ: Wiley.

Dillman, D.A. and M.L. Edwards. 2016. "Designing a Mixed-Mode Survey." In Wolfe, Christof, Joye, Dominique, Smith, Tom W. and Fu, Yang-chih, Sage Handbook of Survey Methodology. Sage Publications Wolf, Joye, Smith and Fu. Thousand Oaks. CA, 255–268.

Duprey, M., J. Murphy, P. Biemer, and R. Chew. 2017. "Veni, Vidi, Vici: Interactive Data Visualizations for Adaptive Total Design." Presented at the 5th Workshop on Adaptive and Responsive Survey Design. Ann Arbor, MI.

Eddy, W.F. and Marton, K., Editors. 2012. *Effective Tracking of Building Energy Use: Improving the Commercial Buildings and Residential Energy Consumption Surveys*. Washington D.C.: The National Academies Press.

Edgar, J., J. Murphy, and M. Keating. 2016. "Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions." *SAGE Open* 6(4): 1–14. Doi: https://doi.org/10.1177/2158244016671770.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R. and S. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society, Series A* 169(3): 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Hardin, M., D. Horn, R. Perez, and L. Williams. 2012. *"Which Chart or Graph is Right for You? Telling Impactful Stories with Data."* Tableau Software. Available at: http://theathenaforum.org/sites/default/files/WHich%20chart%20is%20right%20for%20you.pdf (accessed November 26, 2017).

Hornbaek, K. and E. Frokjaer. 2003. "Reading Patterns and Usability in Visualizations of Electronic Documents." *ACM Transactions on Computer-Human Interaction* 10(2): 119–149. Doi: https://doi.org/10.1145/772047.772050.

Howden, L., S. Joestl, and R. Cohen. 2015. Improving Response Rates using a Mixed-Mode Approach: Results from the National Health Care Interview Survey. Presented at the 2015 FedCASIC Conference. Available at: https://www.census.gov/fedcasic/fc2015/ppt/27_howden.pdf (accessed November 21, 2017).

Kahneman, D. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus, and Giroux.

Laflamme, F. and J. Wagner. 2016. "Responsive and Adaptive Designs." In *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T. Smith, and Y. Fu. Los Angeles: Sage.

Link, M. and A. Mokdad. 2005. "Alternative Modes for Health Surveillance Surveys: an Experiment with Web, Mail, and Telephone." *Epidemiology* 16: 701–704. Doi: 10.1097/01.ede.0000172138.67080.7f.

Luiten, A. and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: an Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society A* 176: 169–189. Doi: https://doi.org/10.1111/j.1467-985X.2012.01080.x.

Morganstein, D.R. and D.A. Marker. 1997. "Continuous Quality Improvement in Statistical Agencies." In *Survey Measurement and Process Quality*, edited by L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. (pp. 475–500). New York: John Wiley & Sons.

Murphy, J., D. Mayclin, A. Richards, and D. Roe. 2016. "A Multi-method Approach to Survey Pretesting." In *2015 FCSM Research Conference Proceedings*. Available at: https://fcsm.sites.usa.gov/files/2016/03/D3_Murphy_2015FCSM.pdf. (accessed November 26, 2017).

Murphy, J., P. Biemer, M. Duprey, and R. Chew. 2017. "Interactive Adaptive Total Design Reports for Near Real-Time Survey Monitoring." Presented at the 2017 Conference of the European Survey Research Association. Lisbon, Portugal.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indictators of Representativeness of Survey Nonresponse." *Survey Methodology* 35: 101–113.

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive Survey Design*. Boca Raton, FL: Chapman and Hall/CRC.

Tufte, E. 2001. *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press. ISBN 0-9613921-4-2.

U.S. Census Bureau. 2015. American Community Survey (ACS) 2014 Data Release New and Noteable. Available at: https://www.census.gov/programs-surveys/acs/news/data-releases/2014/release.html#par_textimage_12. (accessed November 21, 2017).

Zimmer, S., P. Biemer, P. Kott, and C. Berry. 2016. "Testing a Model-Directed, Mixed Mode Protocol in the RECS Pilot Study." In *2015 FCSM Research Conference Proceedings*. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2016/03/G2_Zimmer_2015FCSM.pdf. (accessed November 26, 2017).

Received March 2016
Revised March 2018
Accepted June 2018