Testing a Model-Directed, Mixed Mode Protocol in the RECS Pilot Study

Stephanie Zimmer[1] Paul Biemer[1] Phillip Kott[1] Chip Berry[2]

[1]RTI International, Research Triangle Park, NC

[2]Energy Information Administration, Washington, DC

1. Introduction

The Residential Energy Consumption Survey (RECS) is a periodic survey of households that collects energy characteristics, energy usage patterns, and household demographics. The survey has been conducted since 1978 with the most recent iteration in 2009. Currently, the RECS is an in-person survey using computer-assisted personal interviewing (CAPI) for data collection. The sampling design includes labor-intensive area probability sampling with listing and rostering of housing units along with the use of address-based sampling (ABS) in more recent years in areas that have high coverage of an ABS frame. The Energy Information Administration (EIA) conducted a series of pilot tests to determine the feasibility, cost-effectiveness, time efficiency, and response validity of self-administered modes in the RECS Cities Pilot. The second of these tests, and the subject of this paper, was the RECS Cities Pilot conducted in five localities: Chicago, IL; Jacksonville, FL; San Diego, CA; Seattle, WA; and Worcester, MA. The Cities Pilot study implemented two modes of response – computer assisted web interviewing (CAWI) and paper and pencil interviewing (PAPI). This paper focuses on an experiment to test the so-called “model-directed mode protocol.” This approach attempts to identify households who do not have internet access and would find it very difficult to respond using the web. These households would (initially) be asked to complete a PAPI questionnaire while other households would be requested to respond by web. The experiment described in this paper was designed to determine whether this approach results in a higher response rate than the CAWI-first protocol without reducing the proportion of respondents that complete the questionnaire by web.

2. Sampling Design

The sampling frame for the Cities Pilot was constructed using the U.S. Postal Services’ Computerized Delivery Sequence (CDS). For this study cities are defined using the Census Bureau’s Public Use Microdata Areas (PUMAs). For logistical reasons, drop points were excluded from the sample; however, due to their small number, there should be no consequential loss of generality of the study results.

The sampling frame was implicitly stratified within each city by three variables, in order: socioeconomic status (SES), housing unit size, and internet access propensity. The first two variables, SES and housing unit size, were estimated at the census block group level. Internet access propensity was estimated at the housing unit level. Upon sorting, a systematic sample of 1,071 addresses per city was selected. Within each city, one-quarter of the sample was assigned to each of four treatment combinations defined by interview length and mode assignment protocol. To balance the implicit stratification variables across treatment combinations, the sampled addresses were assigned to treatments in the order in which they were selected. That is, the first sampled address was assigned to the first treatment combination, the second to the second treatment combination, and so forth. Finally, within the two panels assigned to the model-directed protocol, HUs were assigned to either CAWI-first or PAPI-first according to their predicted internet access propensity score. HUs in the lowest quartile of the internet access propensity distribution were assigned to the PAPI-first mode while the remainder was assigned to the CAWI-first mode.

3. Model-Directed Mode Experiment

Internet access in the US is not universal, and even among those with access, many may prefer completing a PAPI survey. We based this assumption on findings in the Home Energy Use Survey (HEUS) (Neufelder, 2014). The

HEUS was the first pilot test of self-administered modes for the RECS and was conducted by the Joint Program in Survey Methodology at the University of Maryland. Although survey literature indicates that offering respondents a mode choice can depress response propensity, it would be impossible for some to respond by CAWI and thus keeping PAPI as an option would seem a viable strategy. Our strategy was to offer PAPI response initially to sample members who are very unlikely to have internet access and thus response by CAWI would to be quite difficult. Thus, these households would have low response propensity to the initial CAWI request.

We designed an experiment to test whether a model-directed mode protocol results in a higher response rate than the CAWI-first protocol without reducing the proportion of respondents that complete the questionnaire by web. Half of the sample, the CAWI-first group, were sent a CAWI invitation first. The other half of the sample, the model-directed group, was assigned an initial mode based on an internet access propensity model score. Sampled households in the first quartile of the internet access propensity distribution received a PAPI questionnaire invitation first, while the remaining households received the CAWI-first invitation. Subsequently, nonrespondents in both groups received a second and, in some cases, third invitation to respond by their choice of either the CAWI or PAPI mode. There was equal allocation of sample to each city and treatment groups. Total sample distributions and sample distributions for each city are shown in **Table 1** for the model-directed experiment, as well as the length experiment.

Table 1: Sample Allocation by Treatment for RECS Cities Pilot

	Total Sampled Addresses	Sampled Addresses per City
Total	5,355	1,071
Short form (~20 minutes)		
Total	2,678	~536
CAWI-first protocol	1,339	~268
Model directed protocol	1,339	~268
Long form (~30-35 minutes)		
Total	2,677	~535
CAWI-first protocol	1,338	~268
Model directed protocol	1,339	~268

4. Propensity Model and Protocol

To predict internet access propensity, we used data from the 2013 American Community Survey (ACS) and marketing data from the Axciom database. Because the model was built on the ACS and applied to the sampling frame, the variables we considered for predicting internet access were necessarily restricted to those that were on both the sampling frame and the ACS. Variables were selected using a classification and regression tree (CART) model. Age of householder, education, and household income were determined to be the only statistically significant variables and therefore these three variables comprised the model used to predict internet access propensity for a given household.

One complication in applying this model was the missing data on the variables age, education, and income using the CDS as a sampling frame. However, for most housing units on the frame, a marketing category is assigned to households which was used to impute the missing variables. Each of these marketing categories groups people of similar age, income, education and the data on the distributions of these three variables within a category, which was used to impute internet access propensity when any one of the variables was missing. For example, one marketing category is called “First Digs” which contains only persons age 24-29. Most in this group have completed high school (78.6%) with 19% completing college, 1.7% completing graduate school and 0.7% attending vocational/technical training. Further, we know that the income of this group ranges from $20,000 to $49,999 with 44.2% with incomes less than $30,000 and 47.9% with incomes between $30,000 and $40,000. Using the ACS model estimates of internet-access propensity by age, education, and income, we calculated a weighted average propensity for persons in the First Digs category using the above distributional information as weighting factors for the missing variable(s). This average propensity was then assigned to persons with missing values on one or more of the three variables. For example, if age is missing but income and education are known, a person in the First Digs category was assigned the average propensity for persons aged 24-29. Finally, the housing unit level internet access propensity computed as the minimum propensity for persons living in the unit.

To model internet access propensity, we partitioned the patterns of missing data into four groups: (1) all three variables are present, (2) one or two variables are present, (3) all three variables are missing but a marketing category is available, and (4) no variables are present and no marketing category is available. Most frame members fall into Patterns 1 and 2 (see **Table 2**). The data available on the frame is at the person level yet our internet access is a household level variable. To address this issue, we estimated internet access propensity for each person and then assigned the minimum propensity of the household members to the housing unit. To deal with the approximately 30% of households in group 4, we used a nearest neighbor imputation strategy. Next we discuss the model used for each of the four patterns/groups.

Table 2: Number of people in each data availability pattern

	Pattern 1	Pattern 2	Pattern 3	Pattern 4
Percentage of People	44.91	49.69	0.61	4.79
Number of People	2,469,450	2,732,349	33,632	263,660

Pattern 1: All Variables Present

If all of the demographic variables were present for an individual, the propensity score model was straightforward. Using the ACS data, we have discretized age, income, and years of education. Each combination of three variables creates a cell in the age by income by education table. Then each person on the Acxiom database was assigned to one cell of this cross-classification table. Finally, the propensity assigned to each person in a cell was based on a logit model for internet access fitted to ACS data using these three variables.

Pattern 2: One or Two Variables Present

If one or two variables were present, we can still use the ACS data to estimate the percentage of people with access to the internet with those demographics. For instance, if only age and income were present (education is missing), we integrated out education to get a propensity estimate for an age by income cell.

Pattern 3: Only Have Marketing Category

Each of the 70 marketing categories describes a unique subset of the United States. For example, First Digs is described as “young, single urbanites who have lower-middle incomes and minimal-to-no net worth.” For each marketing category, the distributions of age, income, and education is provided. Unfortunately, only the marginal distributions for these variables are provided. We assumed that, conditional on the marketing category, age, education, and income are independent.

To estimate internet access propensity using the marketing category, we used a weighted average of the demographics of that marketing category. Let X, Y, and Z define three discrete variables for predicting access propensity having categories i = 1, …, I, j = 1, …, J, and k = 1, …, K, respectively. Let XYZ denote their cross-classification with cells (i,j,k). Let ρ_{ijk} denote the propensity for cell (i, j, k) estimated from the ACS data. Let C denote a marketing category and assume the marginal distributions of X, Y and Z are known for C. Let $p_{Xi}, i = 1, \ldots, I$ denote these proportions for X for the units in C and $p_{Yj}, j = 1, \ldots, J$ and $p_{Zk}, k = 1, \ldots, K$ similarly for Y and Z. Then the propensity was estimated as $\rho_C = \sum_i \sum_j \sum_k \rho_{ijk} p_{Xi} p_{Yj} p_{Zk}$ for category C.

Pattern 4: No Information on Household

Approximately 30% of addresses on the frame had neither education, age, or income data nor marketing categories that could be used for imputing these variables. For these addresses, the internet access propensity was imputed using a nearest neighbor approach. Within each zip code, there are carrier routes indicating the geographic area to which a single mail carrier delivers mail. By definition, housing units within a carrier route should be located close to one another. Additionally, we have the walking sequence within a carrier route which is the order in which the mail carrier goes from unit to unit. Thus, if two housing units have the same zip code and carrier route and their walking sequence number is only one unit apart, they are likely to be very near to each other, perhaps even neighbors in the literal sense.

We took advantage of this information for imputation, trying to first borrow internet access propensities from walking sequence neighbors, then going to the carrier route within zip code level, then the zip code level, and finally the overall level.

5. Results

There was no statistically significant difference in the response rate (p=0.01) between the CAWI-first and model-directed groups. The response rates were 38.02% and 38.72% for the CAWI-first and model-directed groups, respectively. **A logistic** regression model that included indicator variables denoting the standard vs. model directed treatment (denoted by M), the cities, cases above/below bottom quartile cut-off for PAPI-first, and all interactions was fit to predict both response rate and web completion rate. For response rate, none of the interactions were significant so a model with just main effects was fit. Under this model, neither factor M nor the propensity threshold variables were significant. However, for web completion under the reduced model, the interaction of the threshold and the factor was significant after removing interactions involving the cities. This suggests that persons in the model-directed group who completed the questionnaire by PAPI are different than those who used CAWI which is not unexpected.

Table displays the response rates and web completion rates by protocol. There were more web responses in the web-first group (p<0.0001).

A logistic regression model that included indicator variables denoting the standard vs. model directed treatment (denoted by M), the cities, cases above/below bottom quartile cut-off for PAPI-first, and all interactions was fit to predict both response rate and web completion rate. For response rate, none of the interactions were significant so a model with just main effects was fit. Under this model, neither factor M nor the propensity threshold variables were significant. However, for web completion under the reduced model, the interaction of the threshold and the factor was significant after removing interactions involving the cities. This suggests that persons in the model-directed group who completed the questionnaire by PAPI are different than those who used CAWI which is not unexpected.

Table 3: Response rates and web submission rates by protocol and propensity threshold with standard errors in parentheses

Propensity threshold	Web-First		Model-Directed		Total	
	Response Rate	Web Completion	Response Rate	Web Completion	Response Rate	Web Completion

Combined	38.02 (0.95)	20.18 (0.79)	38.72 (0.96)	15.51 (0.71)	38.36 (0.67)	17.85 (0.53)
Above	38.65 (1.10)	21.01 (0.92)	38.24 (1.10)	20.18 (0.91)	38.44 (0.78)	20.60 (0.65)
Below	36.08 (1.89)	17.65 (1.51)	40.16 (1.93)	1.41 (0.47)	38.13 (1.35)	9.51 (0.82)

Note that, in **Table 3**, response rate is defined as the number of completed responses per eligible sampled unit. All sampled units are classified as complete, incomplete, inactive, refusal, UPS undeliverable, or Post Office (PO) undeliverable. A unit is considered *complete* if 7 of 10 key items have valid responses and *incompletes* are units that submitted a questionnaire but it was not complete. *Inactives, Refusals, UPS Undeliverables,* and *PO Undeliverables* each correspond to a particular status code. Additionally, some units are ineligible. For example, RECS only includes primary residences so if a housing unit is determined to be a temporary residence, it is ineligible for the survey. To estimate eligibility of undeliverables, the vacancy rate is used where $e = 1 - [PO\ Vacancy\ Rate] * \frac{Sample\ Size}{PO\ Undeliverables}$ where vacancy rates range from 61% to 74% in the cities. Using these definitions of status of sampled units, the response rate is defined as:

$$Response\ Rate = \frac{Completes}{Completes + Incompletes + Inactives + Refusals + UPS\ Undeliverables + e \times PO\ Undeliverables}.$$

6. Testing the Internet-Access Propensity Model

Among respondents, the internet access in the home was compared to the propensity threshold in **Table** using data from the survey responses, ignoring item nonresponse. Indeed, those above the threshold are more likely to have internet access. This was tested using logistic regression using the threshold indicator for each city and controlling for city in the analysis. For each household on the frame, an internet access propensity was calculated using a model built from ACS data. Cities Pilot respondents were asked about their internet access. Using a logistic regression to predict internet access given the propensity, which ranges from 0.28 to 0.99, the propensity is significant in predicting internet access ($p < 0.0001$).

Table 4: Internet access in the home by internet access propensity threshold

Number of Respondents Above	Number of Respondents Below	Percent Access Above	Percent Access Below	Percent Access Overall
1482	485	90.73%	82.20%	88.65%

7. Comparing CAWI and PAPI data quality

We also compared quality of responses for web versus paper respondents by comparing missing rates for key survey items (see **Table**). To evaluate the differences statistically, Holm-Bonferroni corrected statistical significances were determined at the two-sided 0.1 level (Holm, 1979). Under this correction, the difference with the smallest of the 30 p-values would be marked significant if it was smaller than 0.1/30, the difference with the smallest of the remaining 29 p-values would be marked significant if it was smaller than 0.1/29, and so forth until a difference was not marked significant, at which point the marking stopped. A 0.1 level was used because the Holm-Bonferroni correction tends to be conservative. Because p-values are provided, the reader is free to conduct his or her own tests. 19 out of 30 of the survey items had significantly different rates of missing for web compared to paper.

Table 5: Percent Missing by Mode of Response

Variable	Web Missing (%)	Paper Missing (%)	p-value
Type of dwelling	0.00	1.90	0.000*
Year built	18.69	21.48	0.122
Square footage of dwelling	2.08	15.78	0.000*
Is home heated?	0.00	4.28	0.000*
Source of heating	6.32	12.68	0.000*
Fuel for heating	8.75	8.49	0.844
A/C	0.00	2.19	0.000*
Central A/C	0.36	3.83	0.000*
Window A/C	0.18	3.83	0.000*
Hot water fuel	21.2	19.01	0.227
Own or rent	0.22	0.57	0.207
Number of bedrooms	0.55	2.19	0.001*
Washing machine	0.00	1.71	0.000*
Clothing dryer	0.00	1.81	0.000*
Fuel for dryer	2.58	5.05	0.013
Number of times dryer used per week	0.14	3.44	0.000*
Number of refrigerators	0.00	1.52	0.000*
Internet access	0.11	2.57	0.000*
Number of window panes	8.96	11.60	0.054
Fuel for stove/range	0.00	3.43	0.000*
Fuel for separate cooktop	0.91	10.29	0.000*
Fuel for separate oven	0.00	18.75	0.000*
Number of color TVs	0.00	0.95	0.002*
Number of household members	0.77	1.90	0.026
Past-year income	6.67	7.79	0.334
How is electricity paid?	0.87	3.33	0.000*
How is natural gas paid?	1.74	44.58	0.000*
Age of heating unit	19.44	21.88	0.202
Age of water heater	26.56	26.14	0.834
Age of refrigerator	10.73	11.98	0.384

8. Conclusion

The RECS Cities Pilot experiments were designed primarily to learn how people would respond to an energy survey via the web, and if there was a way to direct more people to the web by targeting to those with internet access. A model was built to predict internet access in the home. Although this model was predictive for internet access, it is not highly predictive of mode choice by the respondent.

EIA and RTI are now fielding the third pilot test of new modes for the RECS. Although we did not use an internet access propensity model to assign initial mode, we are continuing to experiment with mode assignment and contact protocols. For example, rather than simply giving the respondents a choice of mode, one protocol gives respondents a larger incentive if they choose to respond by web rather than mail.

References

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.

Neufelder, J. (2014). *Urban Response Rates and Mode Preference in a Mixed-Mode Survey: Home Energy Use Survey.* Survey Practicum II, Joint Program in Survey Methodology, University of Maryland, College Park, MD.