A COMPARISON AND EVALUATION OF TWO SURVEY DATA COLLECTION METHODOLOGIES: CATI VS. MAIL

Paula Weir, Sherry Beri, Energy Information Administration, Benita O'Colmain, Macro International Inc.
Paula Weir, 1000 Independence Ave. SW, Washington D.C. 20585, pweir@eia.doe.gov

KEY WORDS: Data collection methodologies, Matched-pair design, Cost effectiveness, Response rates, Response time, Edit failures

Abstract

In order to assess and evaluate the relative effectiveness of administering a traditionally mail based attribute frame survey using computer assisted telephone interview (CATI) as the primary data collection mode, EIA conducted a pilot study. The pilot study used a matched pair stratified random sample. Each sampling unit was matched on characteristics thought to be related to the ease of responding by mail versus telephone. Two separate surveys were then conducted and tracked by the primary collection mode, mail or CATI. The results of the pilot are presented and compared for two sets of measures, cost and data quality, where data quality is measured through response rates, response time, and response edit failures. The implications and recommendations for applying the lessons learned to a full survey data collection effort are also described.

Study Design

Due to budget reductions for the 1998 EIA-863 frame survey of approximately 22,000 companies, alternative data collection methods were examined for potential cost savings. Because of the extensive number of follow-up phone calls required for nonresponse and edit failures, EIA considered changing the traditionally mail based survey to a computer assisted telephone interview (CATI) survey. To evaluate the feasibility of CATI as the primary data collection method, and to determine its cost effectiveness, EIA conducted a pilot study designed to allow comparisons between the two data collection methods.

The pilot study used the previous 1994 EIA-863 survey as the sampling frame. Companies who reported an active status at that time and were not reporting on the annual survey were considered to be in scope. Companies selling petroleum products in four or more states were eliminated to simplify the programming for the pilot. This resulted in 20,419 companies that could be sampled for the pilot study. Two variables from the 1994 survey were then used for stratification: 1) number of states reported and, 2) number of items reported. These variables were considered to be directly related to the respondent's initial decision to complete the survey by phone or by mail and to the cost of conducting the survey for either data collection method. The pilot study used a matched pair design with each sampling unit matching on the two stratification variables. The previous survey's respondents, as described above, were then allocated to each stratum for each of the two data collection components: CATI and Mail. A random sample of 500 permitted 95 percent confidence intervals on estimates with plus or minus 5 percent. The majority of companies in the previous survey were one state-companies. Half the sample was therefore allocated to the single state-status-companies and half to the two and three state-status-companies. A fixed number was then allocated to each number-of-items-reported stratum within the number-of-states-reported groupings, except in the three state-status-cells where the population was not large enough to achieve the allotment designated. This yielded a total sample size of 568 for each survey instrument mode, as shown in Table 1, sufficient for the confidence intervals stated above.

Table 1. Pilot Sample Allocations by Stratum

	One State	Two States	Three States	TOTAL
0 Items	50	25	5	80
1 Item	50	25	10	85
2 Items	50	25	35	110
3 Items	50	25	21	96
4 Items	50	25	22	97
5 or more	50	25	25	100
TOTAL	300	150	118	568

The design was intended to allow tests for significant differences between the two instruments for matched pairs, as well as yield overall and survey process cost estimates, and some quality estimates for the two instrument populations. However, it was also intended that the pilot survey reflect the upcoming full survey and survey respondents' behaviors. In that survey, it was decided that respondents would not be forced into one or the other instrument modes. While a respondent would be designated as CATI or mail, the respondent could choose to report either way. A CATI designated respondent would be allowed to report by mail if preferred. Similarly, mail nonrespondents would be followed up via CATI after the form due date. Letters, surveys, and instructions were mailed to both groups, but the CATI group was told that they would be called and their data collected over the telephone. Quality indicators such as response rates, and edit flag counts were then tracked for the two surveys. Total costs and average costs, exclusive of programming costs, were estimated for the two modes for the main processes of the surveys.

Evaluation of Survey Quality Indicators

Response Rates

Response rates were tracked for CATI designated respondents who reported by CATI or by mail, and for mail designated respondents as shown in Figure 1. Final response rates for the two surveys are shown in Table 2, broken down by the number of states and the number of items, excluding respondents who reported by both modes. A total of 518 survey responses were obtained for the CATI component yielding a response rate of 91.2 percent. This response rate was achieved over a 15 week time frame. A total of 500 survey responses were obtained for the Mail component yielding an overall response rate of 87.7 percent. The time to achieve the Mail survey response rate was 23 weeks. These response rates represent the number of forms completed for each of the two designated components, CATI or Mail, regardless of the data reporting/processing mode. The individual sampling cells showed no significant differences between the two components for any of the nine pairs of cells, except one cell. The response rates for respondents who were expected to report only one item, with 96 percent response for the CATI component and 87 percent response for the mail component, were different at the .027 level. The overall response rate was significant just beyond the .05 level, at .056. This finding can be compared to previous studies indicating no difference in response rates (Groves & Nicholls, 1986; Catlin & Ingram, 1988; Groves & Mathiowetz, 1984). A total of 381 responses were obtained through the CATI system and 638 responses were gathered through the mail. Of the CATI responses, 340 were originally selected for the CATI component, and 41 were follow-up responses for the mail component. Of the mail responses, 460 were originally selected for the mail component, while 178 were originally selected to be collected through CATI. Thus, 60 percent of those originally selected to respond through CATI actually did, while 80 percent of those originally selected for mail responded through the mail. The majority of mail responses for the CATI component were received immediately after the mail-out of the survey form, before the start of CATI phone calls.

Figure 1
EIA-863 Pilot Study Response Rates

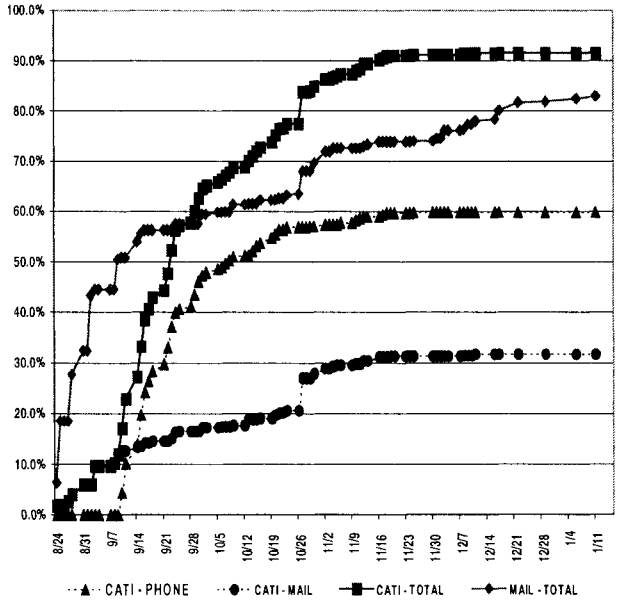

Table 2. Pilot Response Rates

	CATI Component	Mail Component	P-value
TOTAL	**91.20%**	**87.72%**	**0.056**
One State	89.33%	86.33%	0.262
Two States	93.33%	88.67%	0.159
Three States	93.22%	90.00%	0.372
Zero Item	73.75%	66.25%	0.304
One Item	96.47%	87.21%	0.027
Two Items	91.82%	91.82%	1.000
Three Items	95.83%	93.75%	0.519
Four Items	93.81%	92.86%	0.513
Five Items	93.00%	90.00%	0.449

Data Quality Flags

Data were tracked by the number of potential errors that were flagged during the CATI phone survey. Also, the number and type of remaining problems found with the CATI data were tracked by running the data through a batch edit program originally designed for the processing of mail responses. The number and type of problems found after processing all forms received by mail through the batch edit program were also tracked. The survey collected data on companies' operational status and their sales volumes for a set of petroleum products by state. Batch edit failures, therefore, were classified into two types, control data errors and volume data errors, reflecting the different types of data. Control data errors occurred when the responses to the survey questions regarding the company's operational status were not consistent. Volume data errors occurred when a currently reported sales volume was not consistent with a previously reported sales volume, or when reported totals were not consistent with the sum of individual volumes.

Control Data Quality Flags

Comparisons of the control data quality (CDQ) flags were made between the two reporting modes as the respondent reported, regardless of the company's original allocation designation, by the type of edit check failed. Any one company could fail more than one check, and some failures were correlated. The percentages each CDQ type was of the total of all CDQs flagged are shown in Table 3. Even though the number of respondents allocated to each instrument was equal, the number of survey forms processed CATI and mail was not equal, so only the percentage comparisons are appropriate.

From the table it can be seen that, overall, the number of flags set as a percent of forms processed was 0.7 percent less for CATI than mail. This translates into 21.0 percent of the forms processed by mail were identified as failing at least one CDQ edit vs. 18.9 percent of the forms processed by CATI, a 2.1 percent difference in the number of forms. Although the total CDQ

Table 3. Pilot Control Data Quality Flag Distribution

	% Processed Mail	% Processed CATI	Difference (Mail-CATI)
Verifiable CDQs			
CDQ 2	7.00%	4.50%	2.50%
CDQ 13	0.70%	0.00%	0.70%
CDQ 22	0.70%	0.00%	0.70%
CDQ 24	7.60%	19.10%	-11.50%
CDQ 66	43.50%	47.20%	-3.70%
Total of Verifiable CDQs	**59.50%**	**70.80%**	**-11.30%**
Non-verifiableCDQs			
Incomplete/missing status			
CDQ 1	0.70%	9.00%	-8.30%
CDQ 3	0.70%	0.00%	0.70%
CDQ 4	2.80%	0.00%	2.80%
CDQ 6	1.40%	2.30%	-0.90%
CDQ 8	0.70%	0.00%	0.70%
CDQ 9	0.70%	0.00%	0.70%
CDQ 29	7.60%	3.40%	4.20%
CDQ 30	0.00%	1.10%	-1.10%
CDQ 41	1.40%	0.00%	1.40%
CDQ 82	0.70%	0.00%	0.70%
Total incomplete or missing status	**16.70%**	**15.80%**	**0.90%**
Processing conflicts			
CDQ 7	0.70%	0.00%	0.70%
CDQ 28	0.70%	0.00%	0.70%
Manual coding needed			
CDQ 5	20.70%	13.50%	7.20%
CDQ 12	1.40%	0.00%	1.40%
Total processing conflicts	**23.50%**	**13.50%**	**10.00%**
#CDQs/#Forms	**24.10%**	**23.40%**	**0.70%**
# Forms with a CDQ/# Forms	**21.00%**	**18.90%**	**2.10%**

difference was not dramatic, the most basic difference was that there were no flags set for eleven of the CDQs for CATI or, conversely, flags were generated for only eight of nineteen types of CDQs possible in the pilot. Six of these CDQs that resulted in no flags for CATI were non-verifiable CDQs that were the result of missing or incomplete status information reported by the respondent. Three more of these were also non-verifiable, but were the result of changes in the respondent's status that required manual coding or correction of a new error introduced through manual coding. The other two CDQs not experienced by CATI , CDQ 13 and 22, were verifiable flags. It appears that these CDQs, although infrequent for the forms processed by mail, were avoided by the CATI interview due to the skip patterns and edits programmed for the CATI survey. This result can be compared to the findings of Groves and Nicholls, 1986, who found that CATI results in less missing data, as the result of navigation, only in complex questionnaires. In addition, two of the verifiable CDQs (CDQ 24 and CDQ 2) were verified during the CATI interviews. Therefore, these CDQs did not need further research after the CATI interview. When these were removed from the counts of CDQs for CATI, and the percent of processed forms for each component that needed further research in order to clean the data were compared between the two components, it was found that only 14.4 percent of CATI responses needed further work, compared to 21.0 percent of those forms processed by mail.

CDQ 66 (verifiable) constituted the largest percent of the CDQ flags (43.5% and 47.2%) for each type of form processing, but its share of all CDQs was 3.7 percent more for forms processed CATI than for processed mail. This flag, which was set when a company name was entered that was different than the original name, required the interviewer to verify if just the name changed, or if the original company was sold or merged. If the name change was minor and no sale had occurred, the change was verified. If a sale had taken place, or the original company had a completely different name, the original name was put back, the appropriate status response corrected and a new company identification number was issued for the new company name. The status correction then resulted in another CDQ, CDQ5. It was learned from the pilot results that the interviewer screens did not appropriately notify the interviewer to verify name changes, so the scenario described did not occur. Because these name changes can easily be verified within the CATI interview with the appropriate procedures, a large number of flags that would be experienced would easily be verified in the full scale survey if minor adjustments to the pilot interviewer screens were made.

Volume Data Quality Flags

A comparison of the volume data quality (VDQ) flags between the forms processed CATI and mail, regardless of the original designation, is presented in Table 4. Two of the seven VDQ types represented definite errors, non-verifiable (NV), while the other five types were possible errors, verifiable (V), also known as query edits. The percentages in the table were calculated as percent of survey forms processed, not total VDQs, and, therefore, the total percentage can exceed 100 percent. The CATI mode eliminated the definite errors that occurred on 14.7 percent of the mail forms. The verifiable VDQs which occurred the most frequently of the VDQs for both modes were VDQ 3, prior (historical) volume no current volume (57.7% and 49.4% for mail and CATI, respectively), and VDQ 6, current volume no prior volume (43.0% and 43.8% mail and CATI, respectively). Considering both types, verifiable and non-verifiable VDQs, the total percent flagged for mail, 133.4 percent (118.7% plus 14.7%), represents 1.334 volume data quality flags per form processed and for CATI, 119.4 percent represents 1.194 volume data quality flags per form, with the difference in the two modes of 14 percent or .14 flags per form.

Table 4. Volume Data Quality Flags

	Mail				CATI				% Difference
VDQ	# V	% V	# NV	% NV	# V	% V	#NV	% NV	
VDQ 1	N.A.	N.A.	3	0.5%	N.A.	N.A.	0	0.0%	0.5%
VDQ 2	N.A.	N.A.	82	14.2%	N.A.	N.A.	0	0.0%	14.2%
VDQ 3	333	57.7%	N.A.	N.A.	168	49.4%	N.A.	N.A.	8.3%
VDQ 4	40	6.9%	N.A.	N.A.	27	7.9%	N.A.	N.A.	-1.0%
VDQ 5	41	7.1%	N.A.	N.A.	41	12.1%	N.A.	N.A.	-5.0%
VDQ 6	248	43.0%	N.A.	N.A.	149	43.8%	N.A.	N.A.	-0.8%
VDQ 7	23	4.0%	N.A.	N.A.	21	6.2%	N.A.	N.A.	-2.2%
Total	685	**118.7%**	85	**14.7%**	406	**119.4%**	0	0.0%	**14.0%**

Evaluation of Survey Costs by Collection/Processing Mode

The costs attributable to each of the main processes for each data processing mode, regardless of the original designation, were tracked and are shown in Table 5. These figures represent the average cost per form. Programming and management costs, as well as contractor overhead, benefits, General and Administrative, and fees, were not included in this analysis. Costs per form for postage and pre-screening/tracking/filing were fixed. Keying costs per form were a function of the number of states reported. Telephone interview costs for CATI surveys were computed using labor and phone rates per minute multiplied by the total time needed to complete each survey. Batch editing/verifying costs for CATI survey data were fixed based on the average time needed to correct data for a response with at least one CDQ flag. Batch editing/verifying costs for mail surveys were computed based on labor and phone rates per minute times the average time needed to correct data for a form with at least one CDQ, plus the average time needed to correct data for a form with at least one VDQ (computed as a function of the number of states and items per survey form). Average total costs per survey for mail were $1.17 higher than for CATI. Given this savings and estimated initial programming costs for CATI, the break-even point to recover programming investments was roughly 4,300 respondents. This can be compared to the Weeks' rule of thumb (Weeks, 1992) of 1,000 interviews for the break-evenpoint.

Costs were further examined by the number of states and by number of items, using all responses for each data processing mode, regardless of the mode that each response was originally designated. These costs and associated p-values are shown in Table 6. This comparison indicated that CATI data processing costs were not only significantly lower than mail data processing costs for total average cost per survey response, but also for costs within the cells of the two stratification variables, number of states and number of items. The exception was the cell for survey responses which originally had no historical volumes (0 items) where CATI costs were higher than mail, but the difference was insignificant. These companies did not complete the 1994 EIA-863 survey but were identified as in scope for the 1998 survey, and therefore were completing the survey for this first time. It also appeared that the cost for two item-responses greatly increased and then adjusted for three item-responses. While the cost savings was greatest for the three state-respondent-stratum, the cost savings for the item-strata was greatest for the one item-stratum, closely followed by the four item-stratum.

Table 5. Average Costs per Response by Data Processing Mode

	Processed CATI Response	Processed Mail Response
Form/Instructions Mail-out Postage Costs	$0.33	$0.33
Form Return Postage Costs	—	$0.33
Form Prescreening/Tracking/Filing Costs	—	$1.00
Data Entry Costs	—	$1.31
CATI Interview Costs	$4.57	—
Batch Editing/Verifying	$0.30	$3.40
Average Cost Per Survey Response	**$5.20**	**$6.37**

Table 6. Average Costs per Survey Response by Data Processing Mode

	Processed CATI Response	Processed Mail	P-value
Average Cost	**$5.20**	**$6.37**	**0.0001**
One State	$4.80	$5.58	0.0007
Two States	$5.27	$6.69	0.0001
Three States	$6.04	$7.92	0.0001
Zero Items	$5.95	$5.88	0.8841
One Items	$4.24	$5.92	0.0001
Two Items	$5.39	$6.43	0.0287
Three Items	$4.92	$6.36	0.0033
Four Items	$5.00	$6.64	0.0001
Five Items or more	$5.55	$6.66	0.0170

Actual Costs for Each Pilot Designated Mode Component and Full Survey Cost Estimates

Although the average cost per survey response for each data processing mode was different, each pilot component consisted of a proportion of respondents from each originally designated mode. In addition, the relative frequency of the respondents in the population for a stratum is different from that used in the pilot. Therefore, the actual cost of conducting the survey for each pilot component as originally designated, regardless of how processed, was evaluated and the average cost for the full survey using sampling weights was then projected. For the pilot, the total average cost per survey response for the mail component as originally designated was $6.20, while the total average cost per survey for the CATI component as originally designated was $5.67. This cost difference of $1.53 was significant with a p-value of .0002. The projected average cost per response for the full population survey, using CATI as the primary data collection methodology, would also be $5.20 compared to $5.56 using mail as the primary data collection method. It appears, therefore, that the reduction in overall coststhe full population survey if CATI were used as the primary data collection method would be about $0.36 per response, a savings of roughly 6.5 percent. This can be explained by two factors: 1) majority of companies in the full population survey are one state-companies and this group had less cost savings than two and three state-companies; and 2) the proportion of respondents designated CATI and actually responding CATI was only 60 percent, while 31 percent was gathered through the mail. This diminished the overall cost reductions that would have resulted had most respondents reported via CATI. Using these cost savings that reflect the distribution of the population within the strata, and the percentage of CATI designated respondents who choose to report by mail, the break-even point to recover initial programming costs would more realistically occur around 14,000 respondents.

Conclusions

The pilot study was conducted to determine the relative efficiency of collecting data over the phone using Computer Assisted Telephone Interview technology as the primary data collection mode versus collecting data using a paper/pencil mail-based data collection mode. In order to assess and compare the effectiveness of each of the two methods of data collection, quantitative measures were evaluated. The results are summarized as follows:

- For the CATI respondent designated component, overall response rates were higher and the time to achieve the response rate was shorter than that of the mail component.
- The CATI collected/processed data were cleaner; most of the control data quality edits and all of the volume data quality edits were resolved during the phone interview. However, additional data cleaning was needed for some CATI responses after being processed through the batch edit program.
- Average costs for collecting/processing data via CATI were lower than average costs for data collected/processed via mail. Cost reductions increased with the number of states and items per survey form.
- Respondents could not be forced to use either data collection method. A sizeable portion (31% in this case) of the full survey population is likely to report through the mail, even if CATI is designated as the primary data collection mode, thus diminishing the overall cost savings
- Other costs not included, such as programming and CATI setup, reduced these potential cost savings and should be used to determine break-even survey size in deciding on CATI vs. mail. Because of the distribution of the entire survey population within the strata, and the high rate of CATI designated respondents reporting by mail, the break-even point for recovering programming costs was estimated at approximately 14,000 respondents.

Bibliography

Catlin, G.& Ingram, S. (1988) "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Kcentralized Interviewing", Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg (editors) *Telephone Methodology*. New York: Wiley.

Groves, R.M. & Mathiowitz, N.A. (1984) "Computer Assisted Telephone Interviewing: Effects of Interviewers and Respondents", *Public Opinion Quarterly*, vol. 48, p. 356-369.

Groves, R.M. & Nicholls, W.L.II (1986) "The Status of Computer-Assisted Telephone Interviewing: Part II--Data Quality Issues", *Journal of Official Statistics*, no. 2, p. 117-134.

Weeks, M.F. (1992) "Computer-Assisted Survey Information Collection: A review of CASIC Methods and their Implication for Survey Operations", *Journal of Official Statistics*, vol.4, p. 445-466.