Potential Utility of Web Based Data Collection Options

Heather Contrino, NuStats
Sam Echevarria-Cruz, NuStats
Jain Shleymovich, NuStats

Introduction

Current survey programs across the United States are struggling with issues of frame coverage and nonresponse. In an environment of limited resources, the research community is putting a great deal of effort into the investigation of frame, contact, and data collection options. Each traditional data collection mode, regardless of the sample frame, has benefits and limitations in terms of nonresponse levels, measurement error, coverage, and other forms of bias. In recent years, there has been an increased propensity to utilize multi-mode designs as a more respondent friendly approach to increasing overall and subgroup response rates. One data collection mode that has become increasingly employed in the private sector as a supplemental data collection option is the Internet. Internet surveys as a stand alone or as an additional data collection option have gained attention because of four key assumptions: (1) require less time; (2) can have the same level of quality as other methods; (3) less expensive; and (4) less resource intensive to execute.[1] While stand alone email or web based surveys raise serious issues regarding sample representativeness[2], traditional frames that offer an Internet mode as a participation option can provide a useful method for improving participation rates. Data collection options, such as the web, allow people to participate at their leisure and people may appreciate the opportunity to choose their response mode.[3]

Utilizing data collected in the USPS Household Diary Study (HDS), this paper provides an examination of the potential effectiveness of the incorporation of web based data collection options in multi stage surveys. The USPS HDS is a national study that has been conducted since 1987 and utilizes an address frame with both telephone and mail based data collection modes. The first stage of the study collects information on mail behavior and household demographics. The second stage requires the household to record all mail sent and received. Data from the second stage is retrieved via mail back diaries. In 2003, a web based data collection option was incorporated into the first stage of the research design. The survey design of the HDS includes an advance mailing to all sampled households which requests participation in the study and provides the household with the option of completing the first stage of the study via a secure study website. Currently 10 percent of participating households complete the first phase of this study via the web (n=950).

The research design of the HDS and level of web response provides a unique research opportunity. This paper first presents a comparative analysis of the demographic characteristics of first and second stage responders across the phone, mail, and web data collection modes. Using household information and completion data from both the first and second phases, inferences on potential sources and impact of nonresponse will also be examined. These comparative analyses will support increased information for two key research questions:

- What are the demographic characteristics of households most likely to choose web, phone, and mail data collection modes when provided with participation options in stage one of the study?
- Is the level of nonresponse in the second stage of the study (mail back only) different across stage one completion mode groups (web, phone, mail)?

[1] Schonlau, Matthias, Fricker, Ronald D, & Elliot, Marc. (2001). Conducting Research Surveys via E-Mail and the Web.
[2] Dillman, D. A. (2000). Mail and Internet Surveys: The Tailored Design Method. New York: John Wiley & Sons.
[3] Yun, Gi Woong & Trumbo, Craig. (2000). Comparative Response to a Survey Executed by Post, E-mail, & Web Form. *Journal of Computer-Mediated Communication*.

Study Methodology

The USPS Household Diary Study (HDS) study uses a two-stage design in which households are recruited to participate in the diary study in a household interview (Stage 1) and recruited households complete a seven-day diary of mail received and sent (Stage 2). This paper has focused on Stage 1 of the study.

Stage 1. The main function of the household recruitment interview is to recruit households to participate in the diary study. In addition, the interview collects information on household and person demographics, recall of mail sent and received, adoption and use of communication technologies, bill payment behavior and attitudes towards advertising. Households complete the recruitment interview via computer-assisted telephone interviewing (CATI) technology or through a web based instrument with secure personal identification number (PIN) access. The FY 2004 household interview consisted of 8,438 completed interviews with an adult member (age 18 or older) in the household. These respondents represented a cross-section of U.S. households. The household interview contained 155 items and took an average of 23 minutes to administer. The completion rate for the FY 2004 study (defined as the proportion of respondents who completed the diary portion relative to all recruited respondents) was 65.7 percent. This represents an increase from 63.2 percent in 2003. No web data collection option was available in 2003.

Stage 2. Recruited households are sent diaries along with instructions and a toll-free "help" telephone number. The night before the assigned diary week begins, reminder calls are made to each household to confirm receipt of the packet and answer any last-minute questions. If the packet was not received by this time, the address is re-confirmed and the household is assigned a new diary week and re-sent the diary packet. The diary packets contain a Certificate of Appreciation, Instruction Booklet, Diary Instrument, examples of mail markings, and a list of frequently asked questions.

The diary instrument is comprised of two parts:

1. The Question Booklet is color-coded by mail classification (e.g. First-Class Mail received, First-Class Mail sent, Standard, Bulk Rate, Nonprofit, etc.). Information to be collected about each mail classification included: type of mail piece (i.e. envelope, postcard, catalog), receiver zip code, sender zip code, mail classification, mail type, sender type, information about advertising enclosed and receiver reaction or responses to it, and timeliness of the mail piece arrival.
2. Seven answer booklets, each specific to a day of the week. Each booklet was arranged by mail classification and color-coded to correspond to the question sheets.

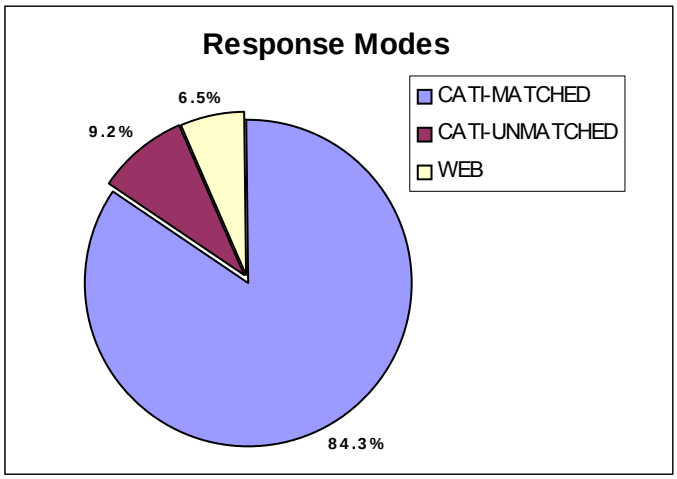
Of the 8,438 households recruited to receive a diary package, 5,541 returned acceptable completed diaries (defined as containing data suitable for analysis) for a completion rate of 65.7 percent.

Response Modes

An advance mailing is sent to all households at the onset of the study. Matched households (households in which a valid telephone number has been identified) are provided with an advance letter and then contacted via telephone to participate in the study. Households where no valid telephone number has been obtained are sent an advance letter as well. This advance letter, however, provides three options for respondents to participate: 1) Return the postage paid card with the households contact information, 2) Call the 1-800 number to participate in the study, or 3) complete the first stage of the study via the Internet. The website address and a secure personal identification number (PIN) are provided to the household at this time. Households selecting options 1 or 2 complete the recruitment interview via CATI.

In Fiscal Year (FY) 2004, 8,438 households were interviewed. Of these households, 7,114 households were contacted directly by using CATI and 1,324 households were given an option to complete the interview via CATI or Web. Figure 1 shows the breakdown of the all three response modes.

Figure 1



Of the total recruited households, 84.3 percent complete via outgoing call, 9.2 percent complete via incoming postcard or call, and 6.5 percent complete via the Internet.

The second stage of the study requires the recruited households to complete daily diaries for an assigned week. A total of 5,541 completed diaries were obtained from recruited households. Note the 5,541 number does not include the diaries that were completed but failed the editing and quality review. Of the 5,541 households, 4,571 diaries were obtained from the matched or CATI only households and 969 diaries came from the unmatched CATI (563) and Web (406) households.

Figure 2

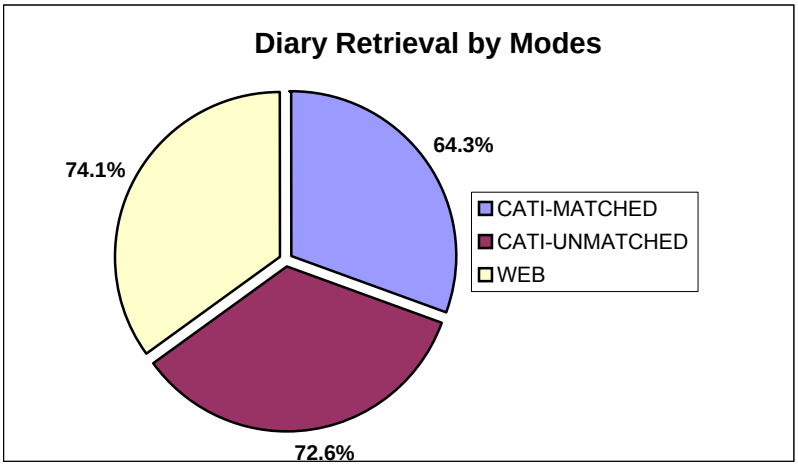


Figure 2 shows the percentage of diaries received by recruitment mode. Overall, unmatched sample performed better by 9.8 percent, where the web retrieval rate was higher by 1.5 percent in comparison to CATI.

Analyses

In order to assess any compositional differences between the study sample by interview source and type, crosstabulations of statistically significant demographic variables were created. Table 1 displays this crosstabulation.

Table 1. Interview Source/Type by Demographics				
	Interview Source/Type			
(18+)	CATI-M	CATI-U	WEB	Total %
Gender				
Male	35.2	32.0	39.5	35.2
Female	64.8	68.0	60.5	64.8
Age				
18-24	2.7	6.6	8.0	3.4
25-34	13.2	15.8	29.0	14.4
35-44	20.1	19.4	26.2	20.4
45-54	22.4	20.7	22.9	22.3
55+	41.6	37.5	13.9	39.4
Race/Ethnicity				
Non-Hispanic White	86.1	79.2	82.8	85.3
Non-Hispanic Black	6.8	10.9	5.9	7.2
Hispanic	3.9	4.4	4.2	4.0
Asian	1.6	1.8	5.7	1.9
Other race	1.6	3.8	1.3	1.7
Martial Status				
Married	65.4	55.8	61.2	64.2
Single/Cohabitating	13.6	21.3	25.5	15.1
Div/Sep/Widowed	21.0	22.9	13.3	20.7
Education				
High School or less	37.1	36.6	13.8	35.5
Some college	22.2	23.6	28.1	22.7
College/Tech Grad.	26.8	25.6	35.3	27.3
Post-Graduate Work	13.8	14.2	22.8	14.4
Employment Status				
Employed full-time	46.6	41.7	68.3	47.5
Employed part-time	11.7	12.2	13.2	11.8
Retired	26.0	24.7	7.2	24.7
Not employed	15.8	21.5	11.3	16.0
Occupation				
Blue Collar	35.5	35.9	13.1	34.2
White collar	47.3	46.1	73.0	48.8
Other	17.2	18.0	13.9	17.0
Income				
<$20K	11.6	14.4	6.0	11.5
$20K--<$35K	14.4	19.7	14.2	14.8
$35K--<$50K	12.5	15.5	10.4	12.6
$50K--<$65K	48.4	41.8	48.4	47.8
Unknown	13.1	8.6	21.0	13.2

PC Ownership				
Yes	78.5	73.7	97.2	35.2
No	21.5	26.3	2.8	64.8
Total N	7,114	776	548	8,438

Various differences are noted within each demographic variable of interest. All variable chi-square tests are statically significant at the $p<0.05$ level. To begin with, males are most likely to be found within the WEB sample at 39.5 percent versus the CATI-Matched (CM) and CATI-Unmatched (CU) at 35.2 percent and 32.0 percent respectively. WEB respondents are twice as likely to be within the 25-34 age range versus all other groups. Conversely, household respondents for both the CM and CU modes are most likely to be within the 55+ age group. Non-Hispanic Blacks are more likely represented within the CU group at 10.9 percent whereas Asians are most likely represented within the WEB respondents. Unmatched respondents are more likely to be single or cohabitating while WEB respondents are less likely to be divorced/widowed/separated. One area where age is not correlated to other demographic variables is within education. WEB respondents, although on average younger, are more highly educated. CATI respondents are 2.5 times more likely to have a high school education or less while WEB respondents are more likely to be both college educated and have experience with post-graduate work. WEB respondents are approximately 44 percent and 66 percent more likely to be employed full-time as compared to CM and CU groups respectively. Conversely, CM and CU respondents are more likely to be retired. The percentage differences between WEB and other respondents increases in reference to occupation, where WEB respondents are more likely to be in the white collar job force. As expected, WEB respondents have almost completely uniform access to a personal computer at home versus CM and CU respondents. Lastly, CU respondents are more likely to lie within lower income categories versus the other two groups.

Table 2 displays the percent distribution of key demographic variables against our main outcome variable of interest, diary completion. We should note that all of the analyses for diary completion as well as recruitment interview was based on respondent demographics (stage1). We however, have no way of knowing if the respondents to stage 1 of the process were the ones completing the diaries.

Table 2. Diary Completion by Demographics

	Diary Completion			
	No	Yes	R%	C%
Interview S/T				
CATI-M	35.7	64.3	100.0	84.3
CATI-U	27.4	72.6	100.0	9.2
WEB	25.9	74.1	100.0	6.5
Gender				
Male	37.7	62.3	100.0	35.2
Female	32.6	67.4	100.0	64.8
Age				
18-24	50.0	50.0	100.0	3.4
25-34	39.4	60.6	100.0	14.4
35-44	35.0	65.0	100.0	20.4
45-54	31.5	68.5	100.0	22.3
55+	32.1	67.9	100.0	39.4
Race/Ethnicity				
Non-Hispanic White	31.2	68.8	100.0	85.3
Non-Hispanic Black	56.0	44.0	100.0	7.2
Hispanic	57.9	42.1	100.0	4.0

Asian	39.0	61.0	100.0	1.9
Other race	39.3	60.7	100.0	1.7
Martial Status				
Married	30.3	69.7	100.0	64.2
Single/Cohabitating	41.5	58.5	100.0	15.1
Div/Sep/Widowed	41.2	58.8	100.0	20.7
Education				
High School or less	40.6	59.4	100.0	35.5
Some college	34.2	65.8	100.0	22.7
College/Tech Grad.	30.8	69.2	100.0	27.3
Post-Graduate Work	25.8	74.2	100.0	14.4
Employment Status				
Employed full-time	35.7	64.3	100.0	47.5
Employed part-time	30.9	69.1	100.0	11.8
Retired	32.4	67.6	100.0	24.7
Not employed	35.9	64.1	100.0	16.0
Occupation				
Not in universe	33.3	66.7	100.0	34.2
White collar	32.4	67.6	100.0	48.8
Other	41.9	58.1	100.0	17.0
Income				
<$20K	44.5	55.5	100.0	11.5
$20K--<$35K	34.7	65.3	100.0	14.8
$35K--<$50K	30.8	69.2	100.0	12.6
$50K--<$65K	31.6	68.4	100.0	47.8
Missing	38.4	61.6	100.0	13.2
PC Ownership				
Yes	32.2	67.8	100.0	79.2
No	42.4	57.6	100.0	20.8
Total N	2,898	5,540		8,438

As we noted in the response rate section of this paper the web respondents are 9.8 percent more likely to complete the diaries then the CM respondents, and CU respondents are 8.3 percent more likely to complete the diaries then CM respondents.

This section examines the demographic characteristics of the respondents who have completed the diary stage of the study. We should note that the respondent's gender does not seem to influence the return of the diaries, 5.1 percent more female respondents have returned the diaries. Age also appears have little influence on diary return with the exception of 18 to 24 year old household respondents. The 18 to 24 year olds run a 50 percent chance of returning the diaries. All other age groups display 60.6 percent to 68.5 percent change of diary return, with 25 to 34 year olds at 60.6 percent and 45 to 54 year olds at 68.5 percent. When examining the race of respondents, we'll see that the highest response rate belongs to non-Hispanic White respondents at 68.8 percent. The Asians and other races complete stage 2, 61.0 and 60.7 percent of the time. In contrast only 42.1 and 44.0 percent of Hispanics and non-Hispanic

Blacks, complete the diaries. The low response by Hispanic group occurs even though they have an option of completing both stages of the survey in Spanish.

Married respondents are more likely to complete the diaries by 11.2 percent then single respondent or the respondent cohabiting, and 8.9 percent more likely to complete the diaries then divorced, separated of widowed respondent. The completion of the diaries increases with the education. Only 59.4 percent of households who have high school or less education will return the diaries, in comparison with 74.2 percent of the post-graduate level households.

When looking at the employment status, the highs return rate can be found in the respondents who work part time (69.1 percent), followed by the retired respondents (67.6 percent) and finally the full-time employed (64.3 percent) similar to unemployed (64.1 percent). The employed individuals tend to be white color employees with 67.6 percent return rate, and they tend to earn $35,000 to $50,000 with 69.2 present return rate. The return rates fall to 55.5 percent with income falling under $20,000 and raise steadily to 68.4 percent for incomes over $50,000. 10.2 present more respondents who return the diaries have at least one personal computer in their household.

Table 3. Logistic Regression of Diary Completion by Demographics

Model 1	B	S.E.	P	O.R.
Interview S/T [CATI-M]				
CATI-U	0.386	0.084	0.000	1.47
WEB	0.464	0.101	0.000	1.59
Constant	0.870	0.043	0.000	---
-2LL	10,815.002			
Nagelkerke R^2	0.007			
Model 2	**B**	**S.E.**	**P**	**O.R.**
Interview S/T [CATI-M]				
CATI-U	0.516	0.089	0.000	1.68
WEB	0.559	0.111	0.000	1.75
Gender [Female]				
Male	-0.364	0.051	0.000	0.70
Age [18-24]				
25-34	0.183	0.146	0.207	1.20
35-44	0.416	0.144	0.004	1.52
45-54	0.649	0.144	0.000	1.91
55+	0.785	0.143	0.000	2.19
Race/Ethnicity [NH-W]				
NH-Black/Hispanic	-0.978	0.075	0.000	0.38
Asian/Other	-0.375	0.128	0.000	0.69
Martial Status [Married]				
Single/Cohabitating	-0.275	0.073	0.000	0.76
Div/Sep/Wid	-0.521	0.065	0.000	0.59
Education [≤ H.S.]				
Some college	0.234	0.066	0.000	1.26
College/Tech Grad.	0.383	0.066	0.000	1.47
Post-Graduate Work	0.581	0.085	0.000	1.79
Income [<$20K]				

$20K--<$35K	0.253	0.093	0.007	1.29
$35K--<$50K	0.284	0.102	0.005	1.33
$50K--<$65K	0.058	0.089	0.513	1.06
Missing	-0.194	0.101	0.055	0.82
PC Ownership [No]				
Yes	0.278	0.067	0.000	1.32
Constant	0.297	0.067		---
-2LL	9,902.339			
Nagelkerke R^2	0.096			

Table 3 displays results of our logistic regression analysis. The goal of this model is to simultaneously account for the effect of demographic variables and interview source/type on diary completion. Logistic regression applies to models where the dependent variable is dichotomous. Instead of slope parameters found within linear regression, exponentiated parameter coefficients represent the odds ratio between the reference category and the indicator category for the independent variable.

The first model (titled "Model 1") displays the regression results for a one-variable model including only interview source/type. With CM being the reference category, we see that CU respondents have 47 percent higher odds of completing their assigned diary while WEB respondents have 59 percent higher odds. The psuedo-R^2 value is small at 0.007% variance explained.

The second model ("Model 2") controls for the effects of the various statistically significant demographic variables. Bivariate relationships from Table 2 are confirmed here. With a set of full controls, male respondent households are less likely to complete diaries along with the non-married and minorities. Education and income are positively related to diary completion. Lastly, the effects of interview source/type remain both strong and significant. In actuality, they increase in magnitude, indicating our average demographic distributions result in lower than possible diary completion rates. The psuedo-R^2 value is higher than model 1 at 0.096 percent variance explained.

Conclusion

In conclusion, both descriptive and inferential analyses identify the factors more likely to influence both participation in the recruitment study as well as diary completion. The inclusion of a web based data collection option in the HDS has increased participation rates for the study by four percent, from 63.2 percent in 2003 to 65.7 percent in 2004. The analysis demonstrates that households selecting the Internet option are more likely to complete the diary. The web option provided a higher overall representation of males, Hispanics, Asians, high income ($50K+), educated, single, employed, and 18-24 year old household respondents in the recruitment stage. The lowest diary completion rates are among males, the non-married, and minorities. Hence, additional representation of these groups in the first stage of the study, as seen with the inclusion of the web option, may have a positive impact on the overall representativeness of the resulting data.

Additional research is needed to examine the impact of multiple data collection modes on the representativeness of the sample. A comparison of web demographics, CATI demographics, and overall demographics for each stage of the study to Census data will provide useful information on how the web option is impacting the representation of specific population groups. Of particular interest for this study is how the inclusion of the web option impacts the representation of population groups that are correlated with mail usage and volume. Certainly an important rational for including the web data collection option for this study was to provide additional representation of population groups most likely to utilize on line bill payment and statement receipt; an important area of revenue loss for the United States Postal Service.