Implementation and Results of the Internet Response Mode for Census 2000

American Association for Public Opinion Research, Montreal, Quebec, May 2001

Erin M. Whitworth, U.S. Census Bureau, Decennial Statistical Studies Division

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Key Words: Internet, Response Mode

The Internet was a viable response option for Census 2000

The United States 2000 Census of Population and Housing offered respondents the first-ever opportunity to respond via the Internet. Preliminary results show that nearly 90,000 attempts were made to obtain an on-line version of the form. This response mode witnessed over 65,000 successful unique submissions. We considered a successful submission one for which we received respondent data. These households had the following characteristics:

- Over 182,000 people were represented from all fifty states, the District of Columbia, and Puerto Rico.

- Average reported respondent[1] age was 42.5.

- Preliminary results show that 78.1 percent of respondents were White, 8.3 percent Black, 4.3 percent Asian or Pacific Islander, and 0.5 percent American Indian or Alaska Native.

- Preliminary results show that 90.6 percent of respondents were not Spanish, Hispanic, or Latino.

- The average time to complete the form was 7.5 minutes.

The Internet response mode presented challenges to the developers and to the Internet Data Collection team. Before Census 2000, we had never used an Internet response mode in a census or census-like test environment. A compressed time frame in which to develop this mode introduced the fear of the unknown regarding system security, integrity, and load capability. Lack of advertising made this option virtually unknown to most people. In spite of these challenges, the Internet Data Collection team successfully implemented an on-line response mode that was accessible, viable, and free of security breaches and operational problems.

This paper will present operational results of this project. It will elaborate on some of the problems we encountered and overcame in the design and implementation of the Internet response mode, discuss preliminary demographic results of the respondents, and address issues of data quality.

BACKGROUND

The Census Bureau had never used the Internet in a previous Census or Census test to collect data.

The Census 2000 Dress Rehearsal in 1998 originally had a plan to include Internet Data Collection, but the plan was abandoned due to security concerns. However, in the fall of 1998, the Department of Commerce decided to reinstate the plan to provide respondents the opportunity to fill out their census forms on-line.

To provide this response mode, the Census Bureau required that the on-line form must securely and accurately collect response data. To create a website to fulfill these requirements, we decided to conduct all aspects of this project in-house, rather than procuring outside contractors which could have increased the cost for this project by hundreds of thousands of dollars (United States Census Bureau, 2001-a).

REQUIREMENTS

The on-site developer had some basic requirements

There was essentially one software developer, and he consented to the project under four conditions. The conditions were:

- Restrict to short form data. The compressed time frame would have made testing and experimentation extremely difficult for a long-form.

---

[1] Respondent refers to the first person listed on the form.

- Require respondents to use their unique 22-digit census ID to access the form. This was necessary to assuage security concerns and prevent the submission of fraudulent forms.

- Provide only an English form. Producing and displaying multilingual forms compatible with myriad web browsers was not possible given the time and resources.

- Provide a point of contact for the developer, a so-called "bureaucracy buffer," to enable him to complete the work, and minimize his time in meetings.

Some of these conditions possibly limited the potential universe of respondents. But in order to safely and quickly provide this response mode, we deemed these requirements necessary.

System Requirements

Usability

The Internet form needed to pass strict usability and cognitive testing. Some preliminary testing occurred at the Census Bureau, and employees filled out sample forms and gave their feedback. We then performed in-depth cognitive testing in which we invited some Census employees to fill out the form while we watched and listened to their reactions to the form.

Respondents answered a series of questions before even getting to the actual form. These questions helped make sure they had the right kind of form, that their browser had the acceptable level of encryption, and that they had their paper form, from which they would retrieve their unique 22-digit census ID to securely fill out the on-line form. We received some good feedback from these test respondents that helped us improve the response process. While we could not change the actual census form, we modified some of the initial questions leading up to the Internet form.

Similarity to paper form

Since we had never tested this response mode, we made every effort to make the on-line version of the form replicate the paper version. We could not change the wording and placement of the questions. This proved to be a challenge since the fact that the form was on-line already made it cognitively different from a paper form.

Ability to handle high volumes

The system needed to handle potentially high volumes of hits and responses. The team simulated different-sized loads and the system passed with flying colors. The system could have handled *substantially* more than the 65,000+ responses we did receive. In fact, it could have handled *millions* of responses. Load testing indicated that we could receive ten million forms per *day*, with the ability to expand the system if necessary. Clearly we did not reach our potential in terms of response.

Security

The system needed to be secure and be impervious to attacks. Requiring the unique 22-digit census ID helped prevent the falsification of data and the "creation" of new households. The unique 22-digit Census IDs are created with two check digits, thus preventing hackers from guessing them simply by random. We did not prevent users from resubmitting the same form ID, but we only processed the first one received. There was still a chance that hackers would resubmit forms under the same ID and disable the system in much the same way that other websites have recently suffered—from overload. The developers anticipated this potential problem and created a safeguard against it. If a respondent repeatedly submitted forms under the same ID, submissions subsequent to the first would be accepted at an increasingly slower rate, effectively removing the respondents' ability to request a new form. These safeguards and preventive measures were effective, and we experienced NO apparent security breaches or attacks on the system.

CHALLENGES

The compressed time frame for development and implementation presented some of the major challenges.

There were challenges to development

As this paper already pointed out, the short amount of time available to develop this response mode presented challenges for the developer. There was simply not enough time to effectively test a long-form version of the questionnaire. Many have questioned the wisdom in this decision, but it was necessary from a development standpoint. We needed to make this mode compatible with a wide range of browsers, and thus we were limited to using mostly HTML programming, and could not use JavaScript, which is unstable in some environments. JavaScript would have enabled us to

have multiple screens for a form, allowing for development of a long form instrument. Putting the long form on one page would cause the form to take an inordinate amount of time to load on the user's screen. Having multiple screens presented some potential security risks that we were unwilling to take, especially as we did not have sufficient time to test alternate methods.

We decided not to include real-time analysis and feedback on the form. For example, if a respondent indicated that there were three people living or staying in the house, apartment, or mobile home on April 1, but then only provided data for two people, we would still accept the form. Another type of feedback would have helped identify discrepancies in the date of birth and age as of April 1. The absence of edit checks was due in part because we had a goal to make the on-line form replicate, as much as possible, the paper form. We did, however, include some post-submission, pre-processing edits. For example, if respondents indicated that the month of birth was "DE," we translated that to a numerical equivalent. Given more time for development, research, and experimentation, we would have incorporated some real-time analysis and feedback.

There were challenges to implementation

A major obstacle in this project was the lack of advertising by design and stated policy. There was no indication anywhere on printed material that one could fill out the form on-line. The Census Bureau did not issue a press release announcing the availability of this response mode. Some have said that unless someone happened to stumble across the link to the on-line form, or had some connection to the Census Bureau, he or she would not have known about this response mode. Clearly we did not receive the number of responses that we had hoped for. The system could have handled hundreds of *millions* of responses instead of the tens of thousands we did receive. It was secure, safe, reliable, and a viable response option.

RESULTS

Preliminary results show that 66,163 households representing 182,748 people used the Internet response mode. These data and all other results reported here are not from final Census files and are preliminary. The data we used to calculate all of the following results come directly from the Internet data capture files from March 3 through April 18. Also, even though we officially closed the Internet response mode at midnight on April 18, one respondent began filling out the form before midnight, but submitted it after midnight. We

kept this form in the universe simply because we received it in time to make the cut for the non-response follow-up operation.

The majority of forms came in long before Census Day (April 1).

Figure 1 shows the total number of Internet forms returned through this mode from March 3 through April 19.

Figure 1. Internet Forms Received Between March 3 and April 19 (Preliminary)
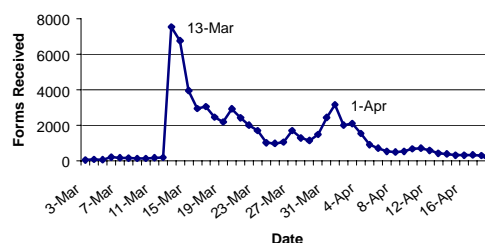
Note that the initial peak occurred on March 13, the date when most households received their census form in the mail. We received a small number of forms before March 13. Some households had already received their form from an enumerator during an operation known as "Update/Leave."[2] There was an additional peak on April 1, "Census Day," which might be because of an assumed April 1 "deadline."

We received forms from every state, the District of Columbia, and Puerto Rico. Puerto Rican households received a census form in the mail in Spanish. We did not provide a Spanish form on-line, but those receiving a Spanish form could submit their responses using an English form. As previously mentioned, we received over 90,000 attempts to respond on-line, but we only received 66,163 successful submissions. Many of the unsuccessful attempts were long-form respondents (who could not use this option), and other invalid attempts to enter in the Census ID. Table 1 shows totals for each region and division.

---

[2] Enumerators in the Update/Leave operation delivered census forms to housing units at the same time they updated the address list. This operation occurred mostly in rural areas.

Table 1. Total Forms Received by Region[3] and Division (*Preliminary*)

	Frequency	Percent	All Modes
Northeast	**9,692**	**14.7**	**19.1**
New England	2,572	3.9	
Mid-Atlantic	7,120	10.8	
South	**26,003**	**39.3**	**36.6**
S. Atlantic	14,272	21.6	
East S. Central	3,593	5.4	
West South	8,138	12.3	
Midwest	**13,422**	**20.3**	**23.3**
East North	9,710	14.7	
West North	3,712	5.6	
West	**16,941**	**25.6**	**21.0**
Mountain	4,594	6.9	
Pacific	12,347	18.7	
Puerto Rico	**105**	**0.2**	
Total	**66,163**	**100.0**	**100.0**

The average length of time to fill out the form was less than the estimated 10 minutes.

The Census Bureau estimated that the average time to complete the *paper* short form would be about ten minutes. Households using the Internet response mode experienced a lower response time, on average. However, it is most likely that households using this mode are not representative of all households.

The amount of time respondents took to complete the form varied quite a bit. The best measure of respondent burden is the time from which the respondent received the form on-screen to the time the respondent submitted the form. This time does not take into account the possibility of the respondent taking a break during the response process, thus increasing the measured elapsed time. The great majority took less than ten minutes; the average was 7.6 minutes. The times ranged from one minute to over 4 ½ hours. It is quite possible that some respondents started filling out the form, and then left it up on their browser while they took a phone call, went to lunch, or took another unrelated break. Figure 2 shows the distribution of the amount of time respondents took to fill out the form.

---

[3] The divisions and regions are as follows:

- The **Northeast** Region includes the New England Division (ME, NH, VT, MA, RI, and CT), and the Mid-Atlantic Division (NY, NJ, and PA).
- The **South** Region includes the South Atlantic Division (DE, MD, DC, VA, WV, NC, SC, GA, and FL), the East South Central Division (KY, TN, AL, MS), and the West South Division (AR, LA, OK, TX).
- The **Midwest** Region includes the East North Division (OH, IN, IL, MI, and WI), and the West North Division (MN, IA, MO, ND, SD, NE, KS).
- The **West** Region includes the Mountain Division (MT, ID, WY, CO, NM, AZ, UT, and NV), and the Pacific Division (WA, OR, CA, AK, and HI).

Figure 2. Amount of time respondent took to fill out the form (*Preliminary*)
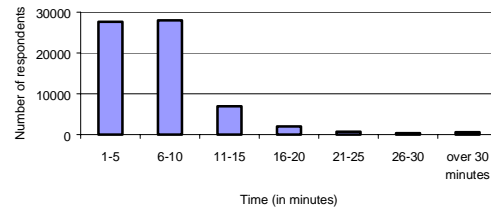
Housing Tenure of respondents

The majority of Internet respondents (65.3 percent) live in owner-occupied housing units. For all response modes, the percent of households living in owner-occupied housing units is 66.0. Table 2 shows the distribution of responses to the question about tenure.

Table 2. Tenure for Internet Responses (*Preliminary*)

	Frequency	Percent	All Modes
Owners	43,224	65.3	66.2
Renters	21,893	33.1	33.8
Missing	1,046	1.6	N/A
Total	66,163	100.0	100.0

These data are as reported, and are not edited. About 1.6 percent of Internet-responding households did not answer the question about tenure or had an invalid response. We have no direct comparison to measure the quality of this rate at a national level for all response modes.

Reported household size

Most households answered the question "How many people were living or staying in this house, apartment, or mobile home on April 1, 2000?" There were only 560 households that either did not answer this question, or provided invalid data. Remember that the Internet form did not have built-in edit checks, which could have safeguarded against this problem. For example, with edit checks, if a respondent entered character data in a numeric field, or left it blank, the Internet form would prompt the respondent to correct the response.

The average reported household size was 2.8, with responses ranging from 0 to 55. Compare this to a national average household size of 2.6 for all response modes. Table 3 shows the distribution of reported household size. Note that only 1.8 percent of households reported household sizes between seven and 12, and 0.05 percent reported sizes larger than 12. The census paper short form, *and* the Internet version, only

had room for data for up to six persons, with a continuation name roster for an additional six. This small percentage of Internet respondents with larger households is an indication that we did not inordinately restrict data collection by limiting the size of the form on-line[4].

Table 3. Reported Household Size *(Preliminary)*

	Frequency	Percent
Missing/Invalid	560	0.8
1	13,430	20.3
2	20,533	31.0
3	12,040	18.2
4	11,569	17.5
5	5,069	7.7
6	1,744	2.6
7-12	1,182	1.8
More than 12	36	0.05
Total	66,163	100.0

Characteristics of respondents.

This paper will only look at responses for the first person listed on the census form. We are looking at raw data responses only.

These preliminary data are all calculated directly from the response files, and are thus raw reported data. Table 4 shows the distribution of reported ages.

Table 4. Reported Age of Respondent *(Preliminary)*

	Frequency	Percent
Missing	1196	1.8
Under 5 years	19	0.03
5 to 9 years	6	0.01
10 to 14 years	8	0.01
15 to 19 years	228	0.34
20 to 24 years	3105	4.69
25 to 34 years	16527	24.98
35 to 44 years	18744	28.33
45 to 54 years	15701	23.73
55 to 59 years	4399	6.65
60 to 64 years	2521	3.81
65 to 74 years	2479	3.75
75 to 84 years	918	1.39
85 years and over	296	0.45
Total	66,163	100.0

The median respondent age was 41.0. The number of households leaving the question blank or providing

[4] The Census Bureau had a telephone follow-up operation to obtain responses for large households.

invalid responses translates to a 1.8 percent item non-response rate.

By and large, the first person listed on the form was reported as male, white, non-Hispanic. Tables 5-7 illustrate this fact. We see that only 1.8 percent of respondents did not respond to the question about sex. For all response modes, there is a slightly different distribution of race (see the column in Table 6 labeled "All Modes"). For all response modes, 12.5 percent of the total population are of Hispanic origin.

Table 5. Reported Sex of Respondent *(Preliminary)*

Sex	Frequency	Percent
Male	49,927	75.5
Female	15,057	22.8
Missing	1,179	1.8
Total	66,163	100.0

Table 6. Reported Race of Respondent *(Preliminary)*

	Frequency	Percent	All Modes
White	51,644	78.1	75.1
Black, African Am., or Negro	5,505	8.3	12.3
American Indian or Alaska Native	338	0.5	0.9
Asian	2,801	4.2	3.6
Native Hawaiian and Other Pacific Islander	65	0.1	0.1
Some Other Race	2,354	3.6	5.5
Two or More Races	1,462	2.2	2.4
No Race Indicated	1,994	3.0	
Total	66,163	100.0	100.0

About 3.0 percent of respondents did not indicate a race.

Table 7. Reported Hispanic Origin of Respondent *(Preliminary)*

	Frequency	Percent	All Modes
Not Spanish/Hispanic/ Latino	59,940	90.6	87.5
Mexican, Mexican Am., Chicano	1,847	2.8	7.3
Puerto Rican	616	0.9	1.2
Cuban	229	0.4	0.4
Other Spanish/ Hispanic/Latino	1,548	2.3	3.6
Two or more boxes marked	106	0.2	-
No boxes marked	1,877	2.9	-
Total	66,163	100.0	100.0

About 2.9 percent of respondents did not answer the question on Hispanic origin.

CONCLUSIONS

The Internet was a viable response mode for Census 2000. Load testing showed that this mode could have handled *hundreds of times* the volume than the over 66,000 responses we did receive. We did not receive any security breaches. For the regular response mode, after the user mailed in the form, we used an elaborate data capture system to convert the responses to data. This Internet response mode had a minimal data capture cost, as the responses were already in data format and ready for processing.

PATH FORWARD

The question is not *if* the Census Bureau will use the Internet for some form of data capture in 2010. The question is *how*. The animal we know as the ubiquitous *Internet* might look the same in 2010, or, and this is more likely, it might look like nothing we currently recognize. Ongoing research will help us determine exactly how to use this response mode in the future.

References

United States Census Bureau, 2001-a, *DMD 2010 Census Planning*, draft.