QUALITY ASSESSMENT OF DATA COLLECTED FROM NON-ENGLISH SPEAKING HOUSEHOLDS IN THE AMERICAN COMMUNITY SURVEY

Pamela D. McGovern and Deborah H. Griffin
U.S. Census Bureau, 4700 Silver Hill Road, Washington, D.C. 20233

Key Words: Item nonresponse; allocation; non-English speaking households

1. Introduction

According to the Census 2000 Supplementary Survey (C2SS), approximately 45 million people aged five years and older spoke a language other than English at home in 2000. Currently, there is little research investigating differences in data quality between English and non-English speaking households. To better understand if differences exist, this paper reports results from a quantitative assessment of data collected from English and non-English speaking households in the American Community Survey (ACS). This research addresses key questions about whether existing methods are resulting in the collection of incomplete data in the ACS due to language barriers.

The ACS, a survey proposed by the Census Bureau to replace the decennial census long form sample, collects social, demographic, economic, and housing data about the nation throughout the decade rather than once every ten years. Data are collected using mail, telephone and personal visit methodologies providing varying degrees of language assistance. It is critical that high quality data be collected for all geographic areas and all population groups. The Census Bureau is interested in developing research strategies and measures of data quality that can be used to assess and improve the quality of demographic survey data obtained from people whose primary language is not English and who have little or no knowledge of English.

This research was undertaken to gain an understanding of which language groups in the United States have the greatest numbers of households with the lowest levels of English proficiency. In addition, the research determined how these households are interviewed in the ACS, and how complete the data collected from these households are. The research focused on non-English speaking households with the lowest levels of English-speaking proficiency because we believe that these households face the greatest challenges in understanding and answering survey questions.

2. Background

The Census 2000 Supplementary Survey and the 2001 Supplementary Survey (SS01) were tests of operational feasibility using the ACS methodology. The supplementary surveys were large-scale surveys of approximately 700,000 addresses each across the United States and were conducted using the procedures and questionnaire planned for use in the full scale ACS.

The surveys were conducted using three modes of data collection to contact households. The first mode uses self-enumeration. The self-enumeration procedure involves the mailing of a pre-notice letter, a survey questionnaire package, and a reminder card. The questionnaire mailing packages include general information about the ACS, and an instruction guide explaining how to complete the questionnaire.

Questionnaires and instruction guides are currently available in English only, but future plans include the development of materials in other languages. The questionnaire provides a telephone number to call if assistance is needed regarding completing the form, including Spanish language assistance. If the original questionnaire is not returned within the specified time frame, a replacement questionnaire package is mailed to the non-responding sample addresses.

Mail questionnaires are checked-in, keyed, and then sent for telephone follow-up if necessary. Telephone follow-up is conducted on cases missing critical information or with household inconsistencies or more than five members in the household. Interviewers located in centralized telephone centers contact these households to obtain all information not present on the mail-returned questionnaire.

For addresses that do not respond by mail and for which a phone number is available, Computer Assisted

---

Telephone Interviewing (CATI) is used to try to reach households in order to complete an interview. The CATI operation is conducted approximately six weeks after the initial questionnaire was mailed. The CATI operation currently is conducted in English and Spanish, but provides no support for those speaking other non-English languages.

Following the CATI operation, a one-in-three sample of the remaining, nonresponding addresses is selected to be sent to the field for Computer Assisted Personal Interviewing (CAPI). Field representatives visit the sub-sampled addresses to try to conduct a personal interview at the nonresponding address. In areas having non-English language needs, interviewers usually are bilingual. CAPI is the last nonresponse follow-up effort.

3. Methodology

3.1 Data Quality Measures

This research was undertaken to assess data quality, focusing on item nonresponse. Item nonresponse occurs when a respondent fails to answer one or more questionnaire items or fails to provide valid responses for questions.

In the ACS, missing data items are compensated for by using imputation procedures. The data from items that were answered are used to impute values for those that are missing or inconsistent. Imputed values can be assigned or allocated. Assignments involve logical imputation where, for example, an answer to another question implies the answer to the missing data item on the same data record. Allocation, on the other hand, involves the use of hot-deck matrices or nearest neighbor households to impute missing data items. Item allocation rates are final measures of completeness that quantify how frequently allocation was the source of data in the production of a specific tabulation. For this reason, we measured item nonresponse by item allocation rates. Allocation rates for questionnaire items are computed as a ratio of the number of housing units or people for which a value for a specific item was allocated to the number of housing units or people for which a response to the item was required.

We calculated item allocation rates by mode of data collection (mail, telephone, and personal visit) for households that speak English only, for households that speak a language other than English, and for households that are considered to be linguistically isolated (LI). A linguistically isolated household is one in which no household member age 14 years or over reports speaking English "very well". All members of a linguistically isolated household are classified as linguistically isolated, including members under age 14 years who may speak only English.

We calculated a combined allocation rate across all population items and across all housing items. The combined allocation rate for all population (housing) items is the ratio of the total number of population (housing) items for which a value was allocated to the total number of population (housing) items for which a response was required. This combined measure was used instead of simply averaging all item allocation rates to ensure that the resulting rate indicated the total amount of required data allocation. If we had simply averaged the item allocation rates, each question would have been given the same weight, regardless of the size of the question's coverage.

3.2 Data and Weighting

This research used data from the C2SS and the SS01 after all edits and allocations had been made. We pooled two years of data and produced two-year averages in order to produce more reliable estimates. The data are weighted to reflect the C2SS and SS01 sample design and include weighting to adjust for noninterviews and coverage errors. We produced standard errors for the allocation rates and compared the rates for non-linguistically isolated and linguistically isolated households to the rates for households speaking English only to detect differences at the 90 percent confidence level.

The estimates in this report are based on responses from a sample of the population. As with all surveys, estimates may vary from the actual values because of sampling variation or other factors.

4. Findings

4.1 Which languages have the greatest numbers of linguistically isolated households?

According to data from the C2SS and SS01, Spanish represents the largest non-English language group in the U.S. with an estimated 10.4 million households of which an estimated 2.7 million are considered to be linguistically isolated. Spanish linguistically isolated households represented 60.8 percent of the estimated 4.2 million linguistically isolated households in the U.S.

Table 1 summarizes results on the number of linguistically isolated households, by household

language[1]. Weighted estimates are provided of the total households reporting speaking each of these languages and the proportion of those that were determined to be linguistically isolated. For example, approximately 26 percent of the households speaking Spanish were considered to be linguistically isolated. The percentage and cumulative percentage of all linguistically isolated households are also provided. The table is ranked by the "percent of total LI households." The top five language groups with an estimated count of 100,000 or more linguistically isolated households are shown in Table 1.

4.2 How were linguistically isolated households interviewed?

Table 2 shows the two-year average distribution of interviews across the three data collection modes (Mail, CATI, and CAPI) for all occupied households in the C2SS and SS01, for those speaking English only, and for households which speak a language other than English. The table shows non-linguistically isolated and linguistically isolated households that fall into each of the five largest linguistically isolated household language groups.

These data show that linguistically isolated households had lower percentages of response by mail than households speaking English only. Spanish linguistically isolated households had an especially low percentage of households who returned the mailout questionnaire, 24.7 percent, and a much higher percentage interviewed in person using CAPI, 62.5 percent.

The mail interview distributions for the non-linguistically isolated households were generally lower relative to the households speaking English only. However, Chinese non-linguistically isolated household actually had a higher response by mail than households that speak English only. Of the non-linguistically isolated households, Spanish had the lowest response by mail, 46.1 percent, but this was 21 percentage points higher than the Spanish linguistically isolated households.

4.3 How complete are the data collected from linguistically isolated households?

Using the C2SS and the SS01 data, we calculated allocation rates to see if there was any evidence that we are collecting less complete data from households with lower levels of English proficiency. The rates were calculated by mode of data collection to determine if mode has an effect on completeness.

Tables 3 and 4 list the combined allocation rates for all housing items and all population items by mode. These summary tables give us an overall picture of the completeness of the data by language group. Significant differences in the mail housing and population allocation rates were found for virtually all five non-English language groups for both linguistically isolated and non-linguistically isolated households when compared to households speaking English only. This result is not surprising given that the questionnaire was available in English only.

The data show that we get more complete data from CATI and CAPI than from mail-returned questionnaires. It is likely that the main reasons why CATI and CAPI data are more complete than mail-returned data is because CATI and CAPI instruments have built-in edits and skip patterns and telephone and field interviewers (who are usually bilingual) ensure that they collect the most complete data possible from respondents.

Though the mail allocation rates for Spanish-speaking households are significantly higher than households speaking English only, Spanish-speaking households interviewed by CAPI had significantly lower allocation rates than households speaking English only. Vietnamese non-linguistically isolated households had some of the highest allocation rates for mail and CATI, especially for the population questions.

Overall, these data show that, while the allocation rates for the linguistically isolated households tend to be higher than households speaking English only, there is no evidence of a dramatic loss in completeness for linguistically isolated households.

5. Limitations

The traditional data quality measures used in this analysis provide a useful, but partial, assessment of data quality. Low item nonresponse rates do not necessarily

[1] Household Language--In households where one or more people (age 5 years old or over) speak a language other than English, the household language assigned to all household members is the non-English language spoken by the first person with a non-English language in the following order: householder, spouse, parent, sibling, child, grandchild, other relative, stepchild, unmarried partner, housemate or roommate, and other nonrelatives. Thus, a person who speaks only English may have a non-English household language assigned to him/her in tabulations of individuals by household language.

ensure good quality data. Other assessments from a qualitative standpoint would be necessary to provide additional insight into the quality of data obtained from households with limited English proficiency. For example, preliminary findings from recent focus groups and cognitive interviews indicate that the way ACS interviews are conducted by Spanish-speaking interviewers and the way in which Spanish-speaking respondents interpret and respond to questions in the ACS Spanish computer-assisted instrument have an impact on data quality (Carrasco, 2002 and Carrasco, 2002).

A question on the ACS questionnaire regarding English-speaking ability was used to determine whether or not a household was linguistically isolated. The level of English proficiency collected by this question is based on people's perceptions of their ability. This opinion-type question has shown high response variance (Singer and Ennis 2002).

6. Conclusions and Next Steps

Spanish is the largest non-English language group in the United States and has the greatest number of linguistically isolated households. The other non-English language groups have far fewer numbers of linguistically isolated households.

The ACS interviews more linguistically isolated households by personal visit. Households with the lowest levels of English proficiency might not return the mail questionnaire because they did not understand it. For these households, it is logical that it would be easier for them to give information to a personal visit interviewer versus trying to navigate through an English questionnaire.

The ACS is successful in obtaining complete data from linguistically isolated households using three modes of data collection. These data show that the overall (when all modes are combined) housing and population allocation rates for linguistically isolated households were only slightly higher than the overall allocation rates for households speaking English only. Future research will include analyzing rates for specific questionnaire items and types of questionnaire items (e.g., check box questions and write-in questions) to better understand which questions had the highest rates of allocation.

In addition, more research is needed to determine how we can improve existing methods, such as telephone follow-up operations and language questionnaire assistance, to achieve more complete data from mail-return questionnaires.

Finally, more research is needed to tap into other dimensions that can have an impact on data quality. These other factors include the extent to which linguistically isolated respondents–especially those responding by mail–understand questions in the survey, and the amount and content of training provided to interviewers for conducting interviews with non-English speaking households.

References

Carrasco, L. (2002). "The American Community Survey (ACS) en Espanol: Results of Cognitive Interviews Using the ACS Spanish Language CAPI Instrument", Internal U.S. Census Report.

Carrasco, L. (2002). "Collecting Data from Spanish Speakers Using the American Community Survey (ACS) CAPI Instrument: Current Practices and Challenges", Internal U.S. Census Report.

Singer, P. and Ennis, S. (2002). "Census 2000 Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview." U.S. Census Bureau, Demographic Statistical Methods Division.

Table 1: Summary of Linguistically Isolated Households by Household Language

Household Language Group	Number of Households: Speaking Listed Language	Linguistically Isolated	% Speaking Language That are LI	% of Total LI Households	Cumulative % of Total LI Households
All occupied households	105,623,930	4,393,921	4.2	--------	--------
English only	86,655,932	0	0.0	0.0	--------
Spanish	10,375,325	2,671,805	25.8	60.8	60.8
Chinese	798,276	291,801	36.6	6.6	67.4
Korean	384,168	139,053	36.2	3.1	70.5
Vietnamese	318,074	137,019	43.1	3.1	73.6
Russian	316,151	136,313	43.1	3.0	76.6

Table 2: Distribution of Interview Completion Modes for English-Speaking and Non-English Speaking Households

Household Language Group	% Mail	% CATI	% CAPI	Total
All occupied households	59.5	9.4	31.1	105,623,930
English Only	61.5	9.5	29.0	86,655,932
Linguistically Isolated				
Spanish	24.7	12.9	62.5	2,671,805
Russian	50.7	7.2	42.3	136,313
Chinese	60.3	4.9	35.0	291,801
Korean	49.9	5.2	45.0	139,053
Vietnamese	56.2	6.8	37.0	137,019
Not Linguistically Isolated				
Spanish	46.1	9.0	45.0	7,703,521
Russian	59.2	9.1	31.8	179,838
Chinese	67.6	5.1	27.3	506,475
Korean	56.1	7.2	36.8	245,115
Vietnamese	54.6	6.7	38.8	181,055

Table 3: Two Year Average Combined Allocation Rates for all Housing Items

Language Spoken	All Modes (%)	Mail (%)	CATI (%)	CAPI(%)
Total	5.24	4.67	5.94	6.13
English Only	5.16	4.53	5.88	6.27
Linguistically Isolated				
Spanish	* 6.20	* 7.93	* 6.47	* 5.40
Russian	* 7.06	* 7.31	* 8.94	6.19
Chinese	* 7.46	* 7.02	6.91	* 8.25
Korean	* 7.67	* 7.80	7.44	7.59
Vietnamese	* 7.56	* 8.33	7.47	6.44
Not Linguistically Isolated				
Spanish	5.25	* 5.07	5.81	* 5.30
Russian	5.34	4.47	5.21	7.10
Chinese	* 5.65	* 4.98	6.84	* 7.16
Korean	* 6.21	* 5.67	6.37	7.07
Vietnamese	* 6.13	* 6.60	7.54	* 5.19

* – Significantly difference from English Only at the α=.10 level.

Table 4: Two Year Average Combined Allocation Rates for all Population Items

Language Spoken	All Modes (%)	Mail (%)	CATI (%)	CAPI (%)
Total	5.89	6.83	4.34	4.71
English Only	5.67	6.37	4.02	4.79
Linguistically Isolated				
Spanish	5.54	* 11.67	3.79	* 4.12
Russian	* 6.96	* 9.53	4.67	4.43
Chinese	* 7.33	* 7.46	5.34	* 7.35
Korean	* 7.89	* 9.09	3.80	* 7.08
Vietnamese	* 7.31	* 9.48	4.53	4.79
Not Linguistically Isolated				
Spanish	* 6.45	* 8.99	* 6.05	* 4.07
Russian	* 6.44	* 7.55	3.82	5.27
Chinese	* 7.28	* 7.19	* 5.52	* 7.70
Korean	* 7.08	* 7.66	* 5.94	* 6.41
Vietnamese	* 9.04	* 11.15	* 10.65	5.73

* – Significantly difference from English Only at the α=.10 level.