# 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. 20233-8700**

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

## 1. Introduction

According to the results from the Census 2000 Supplementary Survey (C2SS), the foreign born population grew by 57 percent since 1990 and approximately 45 million people aged five years and older spoke a language other than English at home. Currently, there is little research investigating differences in data quality between English and non-English speaking households. To better understand the scope and depth of these differences, this paper reports results from a quantitative assessment of differences between English and non-English speaking households in the American Community Survey (ACS) using traditional data quality measures.

The ACS, a survey proposed by the Census Bureau to replace the decennial census long form, will collect social, demographic, economic, and housing data about the nation throughout the decade rather than once every ten years. The ACS will be a monthly survey, sampling approximately 250,000 addresses, and data will be 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 assess the completeness of data collected from non-English speaking households using traditional data quality measures to measure item nonresponse. The research focuses on non-English speaking households with the lowest levels of English-speaking proficiency because we expect that these households face the greatest challenges in understanding and answering survey questions.

While the quantitative measures of data quality provided in this report provide a useful and valuable assessment of data completeness, it is only a partial assessment of data quality. Other assessments from a qualitative standpoint would be necessary to provide additional insight into the quality of data obtained from non-English speaking households. For example, preliminary findings from recent focus groups and cognitive interviews indicate that how ACS interviews are conducted by Spanish-speaking interviewers and the way in which Spanish-speaking respondents interpret and respond to questions on the ACS Spanish questionnaire could potentially lead to errors in the data collected. This result would go largely undetected by item nonresponse analysis and other quantitative analyses of data quality.

## 2. Background

The Census 2000 Supplementary Survey (C2SS) and the 2001 Supplementary Survey (01SS) were tests of operational feasibility of ACS methods. The supplementary surveys were large-scale surveys of approximately 700,000 addresses across the United States and were conducted using the methods and questionnaire planned in the ACS.

The questionnaire collects housing data and socioeconomic and demographic information for up to five residents of a household. If a household has more than five persons, the questionnaire asks the respondent to list their names in the spaces provided and informs them that they may be called to provide additional information regarding these persons.

The survey is conducted using three distinct modes of data collection to contact households. The first mode uses self-enumeration methodology. 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 may include the development of materials in other languages if deemed necessary. The questionnaire does provide a telephone number to call if assistance is needed regarding completing the form, or for Spanish language assistance. If the original questionnaire is not completed within the specified time frame, a replacement questionnaire is mailed again to the non-responding sample addresses.

Mail questionnaires are checked-in, keyed, and then sent for telephone follow-up if necessary. A telephone follow-up operation is conducted on cases with insufficient information or with more than five members in the household. Interviewers contact these households by telephone to obtain missing information and additional information for household members whose data were not listed on the original questionnaire.

For sample addresses that do not respond by mail, Computer Assisted Telephone Interviewing (CATI) is used to try to reach households. The CATI operation is conducted approximately six weeks after the questionnaire was mailed. The CATI operation provides Spanish language assistance, but provides no support for those speaking other non-English languages.

Following the CATI operation, a one-in-three sample is selected from the remaining uninterviewed addresses for Computer Assisted Personal Interviewing (CAPI) . CAPI field representatives visit sub-sampled addresses to conduct a personal interview. In areas having 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 all required questionnaire items or fails to provide valid responses for questions.

In the ACS, missing data items are compensated for by using imputation procedures. That is, the data from the items that were answered are used to impute values for those that are missing. 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. Allocation, on the other hand, involves using statistical procedures, such as hot-deck, nearest neighbor, and regression methods, 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, 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 proper weighting. If we had simply averaged the item allocation rates, each question would have been given the same weight, regardless of the proportion of respondents who were asked to answer the question.

### 3.2 Data and Weighting

This research is based on data from the C2SS and the 01SS. The data used were after all edits and allocations had been made. We pooled two years of data (C2SS and 01SS) to produce more reliable estimates and produced two-year average allocation rates. The data are weighted to reflect the ACS sample design, but do not include weighting to adjust for noninterviews and coverage errors. We produced standard errors for the allocation rates and compared the non-linguistically isolated and the linguistically isolated rates to the rates for households speaking English only to detect differences at the 90 percent confidence level.

## 4. Findings

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

According to data from the C2SS, Spanish represents the largest non-English language group in the U.S. with an estimated 9.2 million households of which an estimated 2.3 million are considered to be linguistically isolated. Spanish linguistically isolated households represented 60 percent of the total estimated number of linguistically isolated households, 3.8 million.

Table 1 summarizes results from the C2SS 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, 25.0 percent of the households speaking Spanish were determined 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.

**Table 1: C2SS -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 | 104,819,002 | 4,197,155 | 4.0 | -------- | -------- |
| English only | 86,154,193 | 0 | 0.0 | 0.0 | --------- |
| Spanish | 10,093,142 | 2,520,729 | 25.0 | 60.0 | 60.0 |
| Chinese | 767,427 | 276,719 | 36.1 | 6.6 | 66.6 |
| Vietnamese | 303,872 | 129,787 | 42.7 | 3.1 | 69.7 |
| Korean | 385,002 | 131,680 | 34.2 | 3.1 | 72.8 |
| Russian | 306,058 | 124,928 | 40.8 | 3.0 | 75.8 |

## 4.2 How were linguistically isolated households interviewed?

Table 2 shows the distribution of interviews across the three data collection modes (Mail, CATI, and CAPI) for all occupied households in the C2SS speaking English only and for non-English speaking households that fall into each of the five household language groups with an estimated 100,000 or more LI households.

[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.

These data show that linguistically isolated households have lower percentages of response by mail than households speaking English only. Spanish linguistically isolated households had an especially low percentage of households interviewed by mail, 26.8 percent, and a much higher percentage interviewed by CAPI, 62.2 percent.

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

| Household Language Group | % Mail | % CATI | % CAPI | Total |
|---|---|---|---|---|
| All occupied households | 60.8 | 8.5 | 30.7 | 104,819,002 |
| English Only | 62.6 | 8.6 | 28.7 | 86,154,193 |
| Spanish | 42.8 | 8.6 | 48.6 | 10,093,142 |
| Chinese | 66.5 | 4.3 | 29.3 | 767,427 |
| Vietnamese | 53.7 | 6.0 | 40.2 | 303,872 |
| Korean | 55.4 | 5.9 | 38.8 | 385,002 |
| Russian | 59.2 | 7.3 | 33.5 | 306,058 |
| Spanish LI | 26.8 | 11.1 | 62.2 | 2,520,729 |
| Chinese LI | 63.2 | 4.5 | 32.4 | 276,719 |
| Vietnamese LI | 53.3 | 5.7 | 41.0 | 129,787 |
| Korean LI | 50.4 | 4.2 | 45.4 | 131,680 |
| Russian LI | 57.9 | 6.7 | 35.5 | 124,928 |

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

Using the C2SS and the 01SS 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 what effect the mode has on completeness.

We chose to calculate two combined allocation rates: one across all housing items and one across all population items. These combined rates give an overall measure of completeness for all housing and population items. We pooled the C2SS and the 01SS data together to produce more reliable estimates and produced two-year average allocation rates.

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 quality of 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 LI and non-LI households when compared to households speaking English only.

The data show that we get more complete data for some items 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 not linguistically 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.

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

| Language Spoken | All Modes (%) | Mail (%) | CATI (%) | CAPI |
|---|---|---|---|---|
| Total | 5.25 | 4.66 | 5.94 | 6.18 |
| English Only | 5.17 | 4.53 | 5.89 | 6.32 |
| **Linguistically Isolated** | | | | |
| Spanish | * 6.15 | * 7.88 | * 6.49 | * 5.38 |
| Russian | * 7.14 | * 7.42 | * 8.87 | 6.28 |
| Chinese | * 7.57 | * 7.15 | 6.84 | * 8.28 |
| Korean | * 7.82 | * 7.87 | 7.03 | * 7.84 |
| Vietnamese | * 7.45 | * 8.20 | 7.27 | 6.41 |
| **Not Linguistically Isolated** | | | | |
| Spanish | 5.24 | * 5.04 | 5.81 | * 5.31 |
| Russian | 5.33 | 4.42 | 5.23 | 7.17 |
| Chinese | * 5.76 | * 5.03 | 6.79 | * 7.39 |
| Korean | * 6.11 | * 5.50 | 6.36 | 7.04 |
| Vietnamese | * 6.24 | * 6.65 | 7.92 | * 5.37 |

* – Significantly difference from English Only at the $\alpha$=.10 level.

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

| Language Spoken | All Modes (%) | Mail (%) | CATI (%) | CAPI (%) |
|---|---|---|---|---|
| Total | 5.87 | 6.80 | 4.33 | 4.71 |
| English Only | 5.66 | 6.35 | 4.01 | 4.81 |
| **Linguistically Isolated** | | | | |
| Spanish | * 5.44 | * 11.57 | 3.93 | * 4.02 |
| Russian | * 6.90 | * 9.54 | 4.56 | 4.39 |
| Chinese | * 7.35 | * 7.55 | 4.90 | * 7.35 |
| Korean | * 7.77 | * 9.07 | 4.30 | * 6.88 |
| Vietnamese | * 7.11 | * 9.39 | 4.28 | 4.66 |
| **Not Linguistically Isolated** | | | | |
| Spanish | * 6.38 | * 8.92 | * 6.01 | * 4.04 |
| Russian | * 6.46 | * 7.51 | 3.99 | 5.34 |
| Chinese | * 7.38 | * 7.18 | * 5.96 | * 7.96 |
| Korean | * 7.17 | * 7.72 | * 6.31 | * 6.50 |
| Vietnamese | * 9.05 | * 11.36 | * 9.50 | 5.91 |

* – Significantly difference from English Only at the $\alpha$=.10 level.

## 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 indicate good quality data. With respect to ACS personal interviews conducted in Spanish, this claim is supported by cognitive research recently conducted by Lorena Carrasco (2002).

A question on the ACS questionnaire regarding English speaking ability is used to determine whether or not a household is 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

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-returned 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 LI 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.

de la Puente, M. and Wobus, P. (1994). "An Item Nonresponse and Log-Linear Analysis of the Spanish Language Forms Availability Test." Paper presented at the Annual Meeting of the American Statistical Association, Toronto, Canada, August 13-18, 1994. Published in the 1994 Proceedings, Section on Survey Methods, pp. 583-588.

de la Puente, M. and Wobus, P. (1995). "Final Report of Results from Item Nonresponse Analysis for the Spanish Language Forms Availability Test." The Statistical Research Division, Census Bureau report.

de la Puente, M. and Gerber, E. (2001). "Translating Demographic Surveys: A Blueprint for Guidelines, Best Practices, and Related Research." Statistical Research Division research document.

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 Divison.