# Item Nonresponse Error for the 100 Percent Data Items on the Census 2000 Long Form Questionnaire[*]

John Chesnut, Decennial Statistical Studies Division, U.S. Bureau of the Census
Washington, D.C. 20233

**Keywords**: item nonresponse, long form, household, census tract

## 1. Introduction

A major source of nonresponse error in surveys is item nonresponse. Item nonresponse occurs in the event that a respondent fails to provide an acceptable response to one or more survey items. Like unit nonresponse, item nonresponse can potentially affect the quality of data for a given study if there are systematic differences between the respondents and the nonrespondents (Lessler and Kalsbeek, 1992). Brick and Kalton (1996) argue that item nonresponse arises because a respondent refuses to answer an item on the grounds that it is too sensitive, does not know the answer to the item, gives an answer that is inconsistent with answers to other items, or because the interviewer fails to ask the question or record the answer. This list of potential sources of item nonresponse implies that the respondent and interviewer directly affect item nonresponse. Other factors may also come into play in influencing item nonresponse. De Leeuw (1999) argues that there are four potential sources of item-nonresponse: the method of data collection (mode), the questionnaire, the interviewer, and the respondent. We conducted research to address the problem of relating household and census tract characteristics affecting item nonresponse for the 100 percent data items on the Census 2000 long form questionnaire. Note that the 100 percent data items are collected on both the short and long form questionnaires. The 100 percent data items include age, sex, race, Hispanic origin, household tenure, and relationship to the householder. This paper presents the results of fitted hierarchical linear models used to examine census tract-level and household-level factors that affected item nonresponse error for the 100 percent data items on the Census 2000 long form questionnaire. Ad-hoc analysis revealed the age item to be difficult for proxy respondents. In addition, the race item proved to be problematic for minority householders. This research may benefit other demographic surveys addressing sources of item nonresponse.

Do any studies provide evidence of specific factors related to these sources of item nonresponse? Studies have shown respondent characteristics and factors such as mode of data collection affect item nonresponse. Respondent characteristics such as age and education level are sometimes good predictors of missing data rates. Studies of these characteristics show that less educated and elderly respondents contribute to higher missing data rates (Colsher and Wallace 1989, Herzog and Rodgers 1989, Dillman 1978, Sudman and Bradburn 1974, Converse 1970, Gergen and Back 1966). Groves (1989) discusses how age and education are likely proxy measures of the encoding and retrieval capabilities of respondents. A diminished capacity to perform either of these cognitive processes would affect the respondent's ability to provide a substantive response. In addition to age and education, other studies have shown that mode of data collection can affect item nonresponse. De Leeuw (1992) demonstrated in a meta-analysis study that telephone and face-to-face surveys fare better than mail surveys in terms of item nonresponse unless the question topic is sensitive.

Results from Census 2000 evaluations provide support for examining other factors in our study beyond those already mentioned. In a study of Census 2000 imputation rates for 100 percent data items (long and short form), Zajac (2003) performed extensive bivariate analysis of imputation rates for various classifications. Of particular interest to us were the results of imputation rates by tenure, by mode of data collection, by check-in date of questionnaire, and by respondent (proxy/household member). The results for tenure showed that households classified as "renter occupied" consistently resulted in higher imputation rates than those classified as "owner occupied" for each of the 100 percent data items. For mode of data collection, self-response resulted in higher imputation rates than face-to-face interview. In addition, data collected from proxies resulted in higher imputation rates than data collected from household members. Finally, the results for check-in date of questionnaires show a general trend of a steady increase in imputation rates for self-response questionnaires for the first few months of the census. In addition, for interviewer questionnaires, we see a general upward trend in imputation rates over time within the period that corresponded to the Nonresponse Followup operation of the census.

Another factor potentially related to item nonresponse is the linguistic isolation status of a household. A linguistically isolated household is defined as a household in which all the members at least 14 years old reported speaking a language other than English and reported not speaking English very well.

[*] This report is released to inform interested parties of research and to encourage discussion.

Linguistic isolation of a household clearly can present a barrier in the administration of a survey possibly leading to problems of item nonresponse. Lestina (2003) found that 4.1 percent of households in the 2000 census were classified as linguistically isolated. In addition, these households were shown to have lower education levels than non-linguistically isolated households, and they were less likely to self-respond than non-linguistically isolated households.

The literature has given some direction about specific factors to include in our study, but other factors may be of interest as well. We suspect household size may affect item nonresponse due to the fact that the larger the household size the greater the response burden. Another factor of interest is householder race. Since two of the 100 percent data items are related to race (race and Hispanic origin), which may be viewed as sensitive questions, a person's race may be related to nonresponse on these items. We know that income and education are positively related, so we suspect that the effect income has on item nonresponse will be similar to that of education - those with higher incomes will have lower rates of missing data.

We have discussed a number of factors that may affect item nonresponse under the presumption that the unit of analysis is the housing unit or the questionnaire. Our proposed analysis involves investigating item nonresponse at the household level as well as the census tract level using a hierarchical linear modeling approach. This approach allows us to relate the properties of households and the properties of tracts in which the households reside. Selected factors already discussed can be aggregated at the tract level and included in our model.

From the literature and our assumptions we have identified factors that may potentially affect item nonresponse error. The purpose of this research paper is to validate some of the findings of previous studies as well as investigate other factors that have yet to be studied in relation to item nonresponse error for the 100 percent data items on the Census long form questionnaire. Household characteristics that we propose investigating include the following: householder age, race, and education; household tenure (rented or owned); total household income; household size; and whether the household is linguistically isolated. Other factors that are worth investigating include the mode of data collection for a household, the date the questionnaire was processed, and whether the respondent was a household member or a proxy. At the tract level, we propose investigating tract level characteristics indicating the minority level, mean income, mean household size, level of linguistically isolated households, and renter level.

## 2. Methods

To meet the research objective of studying item nonresponse for the Census 2000 long form questionnaire, a one percent systematic sample of census collection tracts was drawn from the Census 2000 Sample Census Unedited File (SCUF), i.e. records with no imputation or edits. Note there were a total of 60,462 census collection tracts in Census 2000. Block records were sorted by state, county, and tract which uniquely identify census tracts. A random start was selected from 1 to 100 to select the first tract then every subsequent 100th tract was selected to be in sample. This resulted in a selection of 604 collection tracts. Next the sample housing unit data from the sample of tracts was merged with the Census 2000 Sample Edited Data File (SEDF) to obtain the edited/imputed housing unit and person data for each of the sampled tracts. As a result of the merge, 167,424 households nested within 600 census tracts were included in sample[1]. The primary motive for sampling from the SCUF was to include Census 2000 production data in our analysis (e.g., processing date of a questionnaire).

To study the item nonresponse error in the Census 2000 long form data, we measured a completion rate for the 100 percent data items for each household in the sample. Note that the layout of the long form census questionnaire was such that data were collected for each household member in sequence (i.e. person one, person two, ...). For the purposes of this paper, we label person one the "householder." The age, sex, race, and Hispanic origin data items were collected for all household members. The relationship to householder data item was collected for those not identified as the householder (person two, person three, ...). The housing tenure item was collected under the householder section of the form (person one). An item is defined as completed if no values were imputed or edited for the specific item. An item was imputed or edited if the item was missing or inconsistent with other responses on the questionnaire (Zajac, 2003). We define the household item completion rate as the total number of completed 100 percent data items divided by the total number of 100 percent data items that were offered to a given household. More formally, let us denote the outcome for household *i* in census tract *j* as $Y_{ij}$ = Completed Items/(Household Size x 5 Items).

Given the structure of the data in our sample (households nested within tracts) our method of analysis consisted of applying a two-level hierarchical linear model (cf. Raudenbush and Bryk, 2002). At level 1, the units are households and each household's outcome in

[1] The merge between the sample SCUF data and the SEDF omitted four tracts from the sample containing housing unit data since some of the tracts on the SCUF consisted entirely of housing units that were deleted during the editing process used to create the SEDF.

terms of an item completion rate is represented as a function of a set of individual household characteristics. At level 2, the units are census tracts. To determine the best fitting model, we first fit an unconditional means model to allow us to examine the variation in item completion rates across census tracts and to provide a baseline for comparing more complex models. Next, we examine separately the effects due to the level 2 (census tract) predictors and the level 1 (household) predictors. We conclude by combining both types of predictors into a single model (cf. Singer, 1998). SAS PROC MIXED will be used to fit each of the proposed models by specifying a single model equation with fixed and random effects. The household level and census tract level variables to be included in our analysis are described in Table 1.

**Table 1. Description of the Variables Used in the Study**

| *Variable Name* | *Description* |
|---|---|
| *Household-Level Variables* | |
| Householder Age | Householder's reported age as of April 1, 2000. |
| Householder Race | Householder's reported race dichotomized into two categories (1) white non-Hispanic and (2) nonwhite or white Hispanic. |
| Householder Education | Householder's highest reported level of education: No schooling, 4th grade,..., 12th grade (no diploma), high school graduate, less than 1 year of college, 1 or more years of college (no degree), associates degree, bachelor's degree, master's degree, professional degree, and doctorate degree. |
| Household Tenure | Indicates whether the household is owner or renter occupied. |
| Household Income | The total of all individual incomes reported for each person on the questionnaire (in dollars). |
| Household Size | Number of persons reported living in a household. |
| Mode of Data Collection | Response by mail or by face-to-face interview |
| Check-in Date of Questionnaire | Date for which the questionnaire was processed by the Census Bureau |
| Data Source | Indicates whether the household data was collected from a household member (self response or face-to-face) or via a proxy respondent (face-to-face only). |
| Linguistic Isolation Status | Indicator of whether all occupants within a household 14 and older reported speaking a language other than English and reported they do not speak English very well. |
| *Tract-Level Variables* | |
| Minority Level | Proportion of households within a census tract with a non-white or white Hispanic householder. |
| Mean Household Income | Mean of the household level income measures within a census tract. |
| Mean Household Size | Mean household size within a census tract. |
| Linguistic Isolation Level | Proportion of linguistically isolated households within a census tract. |
| Renter Level | Proportion of renter occupied households within a census tract. |

## 3. Results

### 3.1 Unconditional Means Model

We begin by investigating the between and within census tract variation in mean household item completion rates. To do this, we construct an unconditional means model that serves as a baseline for comparing more complex models later. Let us denote the outcome for household $i$ in census tract $j$ as $Y_{ij}$. At level 1, we express a household outcome (i.e., the proportion of completed items) as the sum of an intercept for a household's tract ($\beta_{0j}$) and a random error ($r_{ij}$) associated with the $i^{th}$ household in the $j^{th}$ census tract:

$$Y_{ij} = \beta_{0j} + r_{ij} \text{ where } r_{ij} \sim N(0, \sigma^2) \qquad (1a)$$

At the tract level, we define as the sum of the overall mean household item completion rate ($\gamma_{00}$) and an error term ($u_{0j}$):

$$\beta_{0j} = \gamma_{00} + u_{0j} \text{ where } u_{0j} \sim N(0, \tau_{00}) \qquad (1b)$$

Given that we have random ($u_{0j}$,$r_{ij}$) and fixed effects ($\gamma_{00}$) included in the model, we used SAS PROC MIXED in fitting the unconditional means model to our data . The estimates for the random effects consist of the variance components, $\tau_{00}$ and $\sigma^2$. We find that the estimated between census tract variance in mean household item completion rates ($\tau_{00}$) equals .000268, and the within tract variability in household item completion rates ($\sigma^2$) equals .009044. Furthermore, both variance components are significantly different from zero. These results indicate that census tracts do differ in mean household item completion rates. In addition, the results confirm that household item completion rates do differ among households within census tracts. Note that the variation among households within census tracts is much greater than the variation among census tracts (almost 34 times more). Finally, we calculate the intra-class correlation coefficient by $\rho = \tau_{00} / (\tau_{00} + \sigma^2)$, which tells us how much of the total variation can be attributed to the between census tract variation. From our estimates of $\tau_{00}$ and $\sigma^2$, the estimated intra-class correlation ($\rho$) equals .026564. In other words, the geographic clustering due to census tract accounts for approximately 3 percent of the total variation in item completion rates.

The unconditional means model includes only one fixed effect which is the intercept, $\gamma_{00}$. This represents the average tract-level household item completion rate. From our results of fitting the model to our data, the estimated intercept $\gamma_{00}$ is .9578.

### 3.2 Including Tract-Level Effects (Level 2)

Similar to the unconditional means model, we can treat the household outcome as the sum of an intercept for a household's tract ($\beta_{0j}$) and a random error ($r_{ij}$), except now we treat $\beta_{0j}$ as a function of level 2 predictors or census tract level effects. The census tract level variables of interest to us are 1) the minority level, 2) mean income, 3) mean household size, 4) linguistic isolation level, and 5) renter level. Note that a correlation analysis revealed that linguistic isolation level and minority level were highly related (*r* = .788). In addition, minority level and mean household size exhibited moderate correlation (*r* = .503). As a result, we included the interaction terms for these pairs of variables in our model. In the previous unconditional means model we had only one fixed effect, the intercept; this model adds seven fixed effects. To facilitate the interpretation of the intercept $\gamma_{00}$, we center each of these variables about their grand mean.

$$Y_{ij} = \beta_{0j} + r_{ij} \text{ where } r_{ij} \sim N(0,\sigma^2) \qquad (2a)$$

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(MINORITY)_j + \gamma_{02}(MEAN\_INCOME)_j \\ &+ \gamma_{03}(MEAN\_HHSIZE)_j + \gamma_{04}(LINGISOL\_LEVEL)_j \\ &+ +\gamma_{05}(RENTERS)_j + \gamma_{06}(MINORITY)_j * (MEAN\_HHSIZE)_j \\ &+ \gamma_{07}(MINORITY)_j * (LINGISOL\_LEVEL)_j + u_{0j} \\ &\text{where } u_{0j} \sim N(0,\tau_{00}) \qquad (2b)\end{aligned}$$

From the results of fitting this model, Table 2 shows the estimates of the $\gamma$ coefficients and their respective standard errors. Note that the $\gamma$ coefficients corresponding to the two interaction terms were not significantly different from zero. This indicates that the minority level and household size do not jointly affect the mean household item completion rate for a given census tract. This result also holds true for the interaction effect of minority level and linguistic isolation. Furthermore, the $\gamma$ coefficient corresponding to the main effect for the level of linguistic isolation was not significantly different from zero. The results shown for linguistic isolation may be due to the scarcity of linguistically isolated households within a given census tract, reducing the effect of this variable on the overall mean tract household completion rate. Reviewing the estimates of the significant $\gamma$ coefficients, we find that the minority concentration within a census tract has the strongest effect on a tract's mean household item completion rate followed by the concentration of renter-occupied housing units contained within a tract. The tract mean household size also proved to be significant.

**Table 2. Effects of Census Tract on Household Item Completion Rate**

| *Effect* | *Coefficient* | *se* | *t-value* |
|---|---|---|---|
| Intercept | .9588*** | .000583 | 1644.43 |
| Minority Level | -.06242*** | .01518 | -4.11 |
| Mean Household Income | $1.685\text{x } 10^{-7***}$ | $<1\text{x}10^{-8}$ | infty |
| Mean Household Size | -.00956*** | .001899 | -5.03 |
| Linguistic Isolation Level | -.02278 | .01608 | -1.42 |
| Renter Level | -.02131*** | .003794 | -5.62 |
| Minority Level x<br>Mean Household Size | .009056 | .009211 | .98 |
| Minority Level x<br>Linguistic Isolation Level | .03312 | .05669 | .58 |

* $p < .10$; ** $p < .05$; *** $p < .001$

The mean household income exhibited a trivially small relationship with mean item nonresponse. These results suggest that a more parsimonious model would exclude the fixed effects of census tract mean household income and linguistic isolation level, it is this reduced model we will apply later to the combined level-1 and level-2 effects model. The intercept, $\gamma_{00}$ = .9588, gives the average tract mean household item completion rate for a census tract of average mean household income, mean household size, minority level, renter level, and linguistic isolation level. Note that this mean is very similar to the overall mean given by the unconditional means model.

To investigate the random effects of this model, we examined the covariance parameter estimates. The results from fitting the model using SAS PROC MIXED give us an estimated $\tau_{00}$ = .000140 and $\sigma^2$= .009045. Comparing these results to those found under the unconditional means model, we observe that the variation within census tract, $\sigma^2$, has not changed. However, we find that the variance component representing variation between census tract has been reduced from $\tau_{00}$ = .000268 to $\tau_{00}$ = .000140. This means that 47.8 percent of the between census tract variation in mean household item completion rates has been explained by including the tract-level predictors (minority level, mean household income, mean household size, linguistic isolation level, and renter level). As we did for the unconditional means model, we calculate the estimated intra-class correlation. We find that $\rho$ = .015242 which means that now only one and a half percent of the total variance is due to geographic clustering after controlling for minority level, mean household income, mean household size, linguistic isolation level, and renter level.

## 3.3 Including Household-Level Effects (Level 1)

We are also interested in modeling the household characteristics as potential predictors of the household item completion rate. We included householder age, race, and education; household tenure; household income (sum of all income values reported in dollars); household size; mode of data collection; check-in date of return; data source (proxy-interviewer mode only/household member); and linguistic isolation. Again, we denote the outcome for household *i* in census tract *j* as $Y_{ij}$. We represent this outcome as a function of the individual household characteristics, mode of data collection, check-in date of return, and data source with a model error term ~ $N(0, \sigma^2)$:

$$\begin{aligned}Y_{ij} &= \beta_{0j} + \beta_{1j}(AGE)_{ij} + \beta_{2j}(HHEDUCATION)_{ij} \qquad (3a) \\ &+ \beta_{3j}(HHINCOME)_{ij} + \beta_{4j}(DATE)_{ij} + \beta_{5j}(HHSIZE)_{ij} \\ &+ \beta_{6j}(TENURE)_{ij} + \beta_{7j}(MINORITY)_{ij} + \beta_{8j}(MODE)_{ij} \\ &+ \beta_{9j}(PROXY)_{ij} + \beta_{10j}(LINGISOL)_{ij} + r_{ij}\end{aligned}$$

$$\beta_{qj} = \gamma_{q0} + u_{qj} \text{ where } \mathbf{u}_j \sim N(\mathbf{0},\mathbf{T}),\ q = 0,1,...,10 \qquad (3b)$$

Comparing this model to the unconditional means model, we have added ten household level fixed effects (the $\gamma_{q0}$ terms) and for each fixed effect we have included an additional random effect ($u_{qj}$). Interpreting the meaning of this model, we are now hypothesizing that a household's item completion rate is related to the given household level characteristics and that the relationships to these variables vary across census tract. Specifically, the regression coefficients $\beta_{qj}$, $q = 0,...,10$ indicate how the outcome is distributed in census tract $j$ as a function of the measured household characteristics. Fitting this model to data allows us to assess the strength of the household predictors and to determine the proportion of the total variance among households within census tract these factors explain.

Using SAS PROC MIXED, we found that the computing resources available were not capable of fitting the described model to our data due to the complex nature of the variance-covariance matrix, **T**, given the large number of specified random effects (more than 4,000 across the census tracts). As a result, we were forced to fit a reduced model by reducing the number of random effects in our model. To do so, we chose the top five household properties varying across census tract to be assigned a random slope coefficient ($\beta_{qj} = \gamma_{q0}+u_{qj}$, $q=5,\ldots,9$): household size, tenure, minority status, mode of data collection, and source of data. The remaining variables household income, householder age, householder education, return date of questionnaire, and linguistic isolation status were assumed to have fixed sloped coefficients ($\beta_{qj} = \gamma_{q0}$, $q=1,\ldots,4, 10$). All variables included in the model were group mean centered to facilitate interpretation of the intercept.

Table 3 shows the results of fitting this model to our data. First, we review the results of the estimated coefficients for the fixed effect terms in our model. Note that for each fixed effect the coefficient was significantly different from zero. The average census tract mean household item completion rate, $\gamma_{00}$, was estimated as .9577. The average gap for renter occupied housing units was estimated as -.00557. That is, for a typical census tract, renter occupied households produced a household item completion rate that was, on average, about 6/10 of a percentage point lower than owner occupied households with similar household-level characteristics. The average minority gap was estimated as -.02708, households with a minority householder produced a household item completion rate on average 2.7 percentage points lower than households with a non-minority householder. Similarly, households where the data were collected through an interviewer as opposed to self-response resulted in a item completion rate that was estimated to be on average 1.4 percentage points higher. Households where the data was collected via proxy resulted in an item completion rate that was estimated to be about 13 percentage points lower than data collected from a household member for a typical census tract. Linguistically isolated households resulted in an item completion rate that was approximately 6/10 of a percentage point lower than non-linguistically isolated households. Furthermore, the householder education level and household income are positively related to item completeness. In addition, the householder's age and the return date of the questionnaire are negatively related to item completeness. The magnitude of the relationship for household income was extremely small which may suggest excluding this variable from the combined tract and household level effects model later.

**Table 3. Effect of Household Characteristics on Household Item Completion Rate**

| *Fixed Effect* | *Coefficient* | *se* | *t-value* |
|---|---|---|---|
| Average Census Tract Mean Household Completion Rate, $\gamma_{00}$ | .9577*** | .000713 | 1343.94 |
| Householder Age, $\gamma_{10}$ | -.00041*** | .000016 | -25.94 |
| Householder Education, $\gamma_{20}$ | .002704*** | .000090 | 30.10 |
| Household Income, $\gamma_{30}$ | $1.115\text{x}10^{-8***}$ | $<1\text{x}10^{-8}$ | infty |
| Questionnaire Return Date, $\gamma_{40}$ | -.00017*** | .000013 | -12.88 |
| Household Size, $\gamma_{50}$ | -.00503*** | .000281 | -17.88 |
| Household Tenure, $\gamma_{60}$ | -.00557*** | .000662 | -8.42 |
| Minority Householder Status, $\gamma_{70}$ | -.02708*** | .002088 | -12.97 |
| Mode of Data Collection, $\gamma_{80}$ | .01366*** | .001307 | 10.45 |
| Proxy Data Collection, $\gamma_{90}$ | -.1298*** | .003284 | -39.52 |
| Linguistic Isolation Status, $\gamma_{10\,0}$ | -.00642*** | .001372 | -4.68 |
| *Random Effect* | *Variance Component* | | |
| Census Tract Mean Household Completion Rate, $u_{0j}$ | .000263*** | .000018 | 14.58 |
| Household Size, $u_{5j}$ | .000026*** | $2.477\text{x}10^{-6}$ | 10.46 |
| Tenure, $u_{6j}$ | .000046*** | .000013 | 3.55 |
| Householder Minority Status, $u_{7j}$ | .000554*** | .000130 | 4.26 |
| Mode of Data Collection, $u_{8j}$ | .000357*** | .000033 | 10.75 |
| Proxy Data Collection, $u_{9j}$ | .004147*** | .000390 | 10.64 |
| Level-1 Effect, $r_{ij}$ | .008289*** | .000029 | 286.39 |

* $p < .10$; ** $p < .05$; *** $p < .001$

Focusing now on the random effects, Table 3 shows the estimated variances of the random effects at the household and census tract level ($\sigma^2$ and $\tau_{qq}$). Referring back to the unconditional means model, we see that the within tract variation among households has been reduced from .009044 to .008289 after controlling for the nine household characteristics. In other words, only 8.3 percent of the variation among households within census tract is explained by this household level model.

The estimated variances of the random effects at the census tract level ($u_{qj}$) in Table 3 gives us the estimated magnitude of the variability for these effects across census tracts. We observe that the variability in the intercept and slope coefficients is extremely small across census tracts, ranging from .000026 to .004147. Still, univariate tests of hypothesis showed that the variances for each of the random effects were significantly different from zero. Note that the between tract variance for the tract mean household completion rate $u_{0j}$ = .000263 is approximately equal to the between tract variance found in the unconditional means model (1) $u_{0j}$ = .000268.

## 3.4 Including Both Census Tract and Household-Level Effects (Levels 1 and 2)

Now that we have separately fit models with either census tract-level effects or household-level effects only, we are ready to fit a combined model. Based on the previous results, we decide to omit the tract-level variables mean household income and linguistic isolation level, and the household level variable household income. Thus, we have the following model equation:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(AGE)_{ij} + \beta_{2j}(HHEDUCATION)_{ij} \\ &+ \beta_{3j}(DATE)_{ij} + \beta_{4j}(HHSIZE)_{ij} \\ &+ \beta_{5j}(TENURE)_{ij} + \beta_{6j}(MINORITY)_{ij} + \beta_{7j}(MODE)_{ij} \\ &+ \beta_{8j}(PROXY)_{ij} + \beta_{9j}(LINGISOL)_{ij} + r_{ij} \end{aligned} \qquad (4a)$$

$$\begin{aligned} \beta_{qj} &= \gamma_{q0} + \gamma_{q1}(MINORITY)_j \\ &+ \gamma_{q2}(MEAN\_HHSIZE)_j + \gamma_{q3}(RENTERS)_j + u_{qj} \\ q &= 0,4,....,8 \end{aligned} \qquad (4b)$$

$$\beta_{qj} = \gamma_{qj} \;\; q = 1,2,3,9 \qquad (4c)$$

The goal of this combined model was to investigate how differences among census tracts might influence the effects of household characteristics on item completeness within a census tract. Table 4 shows the fixed effect results of fitting the combined model equation (4). We summarize our conclusions as follows:

- Census tract mean household item completion rate. The minority level of a tract was negatively related with the census tract's mean household item completion rate ($\gamma_{01}$ = -.1090). In other words, mean household item completion rates were lower in tracts with higher minority concentrations. Similarly, mean household item completion rates were lower in Census tracts with higher concentrations of renters ($\gamma_{03}$ = -.01242) and large households ($\gamma_{04}$ = -.03913).
- Household size. The effect of household size within a census tract exhibited a dependence on the census tract's minority level ($\gamma_{41}$ = -.00725), mean household size ($\gamma_{42}$ = -.00303), and renter level ($\gamma_{43}$ = -.00650). Thus, higher values of any of these tract level characteristics strengthened the negative effect of household size on item completion rates within a tract (i.e., after controlling for the remaining household level effects).
- Tenure. The gap in item completion rates between renter and owner occupied households within a census tract was dependent on mean household size ($\gamma_{52}$ = .004958). Oddly enough, the tenure gap within a census tract was attenuated by higher values of mean household size. Note that this conclusion was not highly significant ($p$ < .05).
- Minority Status of Householder. The minority status of a householder within a census tract exhibited a dependence on the minority level ($\gamma_{61}$ = .05785). The minority gap was reduced for census tracts with higher proportions of minority householders. Note that this conclusion was not highly significant ($p$ = .04).
- Mode of Data Collection. The effect due to mode of data collection within a census tract was shown to be dependent on the minority level ($\gamma_{71}$ = .09698), mean household size ($\gamma_{72}$ = .006405), and renter level ($\gamma_{73}$ = .02283). These coefficients indicate a further increase in household item completion rates due to face-to-face interviews for larger values of mean household size, minority level, and renter level.

**Table 4. Interaction Effects for the Combined Census Tract and Household Level Model (significant effects only)**

| *Fixed Effect* | *Coefficient* | *se* | *t-value* |
|---|---|---|---|
| Census Tract Mean Household Item Completion Rate | | | |
| BASE, $\gamma_{00}$ | .9585$^{***}$ | .000559 | 1714.38 |
| MINORITY LEVEL, $\gamma_{01}$ | -.1083$^{***}$ | .01282 | -8.44 |
| MEAN HHSIZE, $\gamma_{02}$ | -.01116$^{***}$ | .002261 | -4.94 |
| RENTER LEVEL, $\gamma_{03}$ | -.03681$^{***}$ | .004205 | -8.75 |
| Household Size | | | |
| BASE, $\gamma_{40}$ | -.00464$^{***}$ | .000268 | -17.32 |
| MINORITY LEVEL, $\gamma_{41}$ | -.00747$^{*}$ | .004308 | -1.73 |
| MEAN HHSIZE, $\gamma_{42}$ | -.00296$^{***}$ | .000806 | -3.67 |
| RENTER LEVEL, $\gamma_{43}$ | -.00658$^{***}$ | .001565 | -4.20 |
| Tenure | | | |
| BASE, $\gamma_{50}$ | -.00578$^{***}$ | .000664 | -8.70 |
| MEAN HHSIZE, $\gamma_{52}$ | .004397$^{**}$ | .002198 | 2.00 |
| Minority Status of Householder | | | |
| BASE, $\gamma_{60}$ | -.02949$^{***}$ | .002470 | -11.94 |
| MINORITY LEVEL, $\gamma_{61}$ | .05804$^{**}$ | .02803 | 2.07 |
| Mode of Data Collection | | | |
| BASE, $\gamma_{70}$ | .01226$^{***}$ | .001215 | 10.09 |
| MINORITY LEVEL, $\gamma_{71}$ | .09762$^{***}$ | .01451 | 6.73 |
| MEAN HHSIZE, $\gamma_{72}$ | .005551$^{**}$ | .002694 | 2.06 |
| RENTER LEVEL, $\gamma_{73}$ | .02110$^{***}$ | .005016 | 4.21 |

$^{*}$ $p$ < .10; $^{**}$ $p$ < .05; $^{***}$ $p$ < .001

Having discussed the fixed effects for the combined model, we now focus on the random effects ($u_{qj}$). Our random coefficient equations (4b) included tract-level predictors to explain some of the variation in the random coefficients across census tracts whereas in our previous random coefficients model (3), these predictors were omitted (i.e., $\beta_{qj} = \gamma_{q0}+u_{qj}$). In other words, the random effect $u_{qj}$ in our combined model is the residual census tract effect unexplained by the tract level predictors minority level, mean household size, and renter level. Therefore, $\tau_{qq}$ is now a conditional variance. Table 5 gives us the estimates of the conditional variances ($\tau_{qq}$) for each of our random effects ($u_{qj}$); for comparison the unconditional variances from model (3) are also given. First, we notice that the variation across census tract in our intercept (i.e., census tract mean household item completion rate) has been reduced from .000255 to .000115. Therefore, as indicated by Table 4, 54.90 percent of the variation in our intercept has been explained by census tract minority level, mean household size, and renter level. This finding is similar to the reduction in variance of the intercept produced in fitting model equation (2). Similarly, we find that the variation in the slope coefficient for household size has been reduced from .000023 to .000018. This implies that 21.74 percent of the variation for this slope coefficient has been explained. Furthermore, the variance in the slope coefficient for mode of data collection has been reduced from .000336 to .000169 resulting in 49.70 percent explanation of the variance. Variation among census tract for the slope coefficients corresponding to tenure, household minority status, and proxy data collection are not explained by the given tract level variables.

**Table 5. Proportion of Variance Explained by Final Model**

| *Random Effect* | *Unconditional Model Variance Component* | *Conditional Model Variance Component* | *Proportion of Variance Explained* |
|---|---|---|---|
| Census Tract Mean Household Item Completion Rate, $u_{0j}$ | .000263 | .000145 | 44.87% |
| Household Size, $u_{4j}$ | .000026 | .000021 | 19.23% |
| Tenure, $u_{5j}$ | .000046 | .000044 | 4.35% |
| Householder Minority Status, $u_{6j}$ | .000554 | .000584 | -5.42% |
| Mode of Data Collection, $u_{7j}$ | .000357 | .000215 | 39.78% |
| Proxy Data Collection, $u_{8j}$ | .004147 | .004172 | -0.60% |

## 4. Conclusion

In summary, we were able to fit in sequence an unconditional means model and separate conditional means models including either household level effects or tract level effects to arrive at a combined model with both household and tract level effects. Fitting these models to our data allowed us to assess the strength of association of census tract and household level factors with household item completeness for the 100 percent data items on the Census 2000 long form questionnaire. In addition, we were able to determine the proportion of the between tract variance and within tract variance these factors explain.

From the unconditional means model (1), we established a small geographic clustering effect on tract mean household completion rates due to census tract geography. However, from the conditional means model (2) and the combined model (4), we found that much of the variation between census tracts in mean household completion rates reflects the concentration of minority householders, renter occupied households, and large households.

Compared to the between census tract variance, the within tract variance among households was much larger and more difficult to explain based on our household-level factors. Controlling for householder age, education, minority status; household tenure; household income; return date of questionnaire; mode of data collection; and source of data, we were only able to explain 8.1 percent of the variation in household item completion rates among households within census tract. This result is somewhat surprising given that we have provided a fairly comprehensive set of household level variables.

Having explained some of the variation in household item completion rates between and within census tract, we were also able to determine the strength of the relationship of household and tract level characteristics with household item completion rates. From our tract level effects model (2) and the combined model (4), we found that census tracts with higher concentrations of minority householders, renter occupied households, or large households pushed completion rates lower. Higher concentrations of minority households within a tract produced a far greater impact on the tract mean household item completion rate than higher levels of renters and large households. We speculate that the effect of minority households on census tract mean item completion rates may be due to a minority household's lack of response to the race and Hispanic origin items since both of these items can be viewed as potentially sensitive questions.

At the household level (model 3), the most noticeable effect on household item completion rates was due to the source of data collection - proxy or household member. Households where the data was collected via a proxy resulted in a household item completion rate that was on average 13.1 percentage points lower than a household where the data was collected from a household member (assuming all other variables are the same). Other effects worth noting included householder minority status and mode of data collection. Lower rates of completeness were attained for households with minority householders. Higher rates were attained for households where their data was collected in a face-to-face interview.

Focusing on how census tract characteristics interacted with household-level characteristics to influence completion rates, we found from the results of the combined model (4) that the lower completion rates produced by large households were slightly increased in tracts with higher levels of minority and renters, and higher mean household sizes. Furthermore, the householder minority gap in item completion rates was reduced for census tracts with higher concentrations of minority householders. Finally, face-to-face interviewing led to higher completion rates, but were even further increased in tracts with higher concentrations of minority householders, renter occupied households, and large households.

To satisfy our suspicion of whether the race and Hispanic origin items were problematic for minority households, we investigated the association of our household-level characteristics with the household completion rate for each of the individual 100 percent data items. Using a random intercepts only model (i.e., $\beta_{0j} = \gamma_{00}+u_{0j}$ and $\beta_{qj} = \gamma_{q0}$, $q = 1,\ldots,10$) we fit a separate model for each item completion rate (tenure, relationship to householder, sex, age, race, and Hispanic origin). Note that this analysis was not part of our original planned analysis and is not included in the results section. Our most notable results from fitting each of the models showed that households where the data was collected via a proxy respondent resulted in a household item completion rate for the age item that was on average 45.1 percentage points ($p < .001$) lower than households where the data was collected from a household member. This implies that the age item is difficult for a proxy respondent to provide an answer. This may be due to the fact that age is not an observable trait or there may be a sensitivity issue associated with providing someone's age without their permission. Furthermore, we found that the completion rate for the race question was, on average, 8.2 percentage points ($p < .001$) lower for households with a minority householder than households with a non-minority householder. This result proves that minority households did have problems responding to the race item. Again, we speculate this may be due to the sensitive nature of the item.

Our research has given us some insight into how household factors and census tract-level factors affect item nonresponse. However, in reality the response mechanism that motivates a respondent to complete a questionnaire item is likely to depend on many factors beyond those that we have covered in this paper. Other factors of interest that may have provided further explanatory power include respondent attitudes towards individual items, survey environment, household composition, household social-demographic characteristics, interviewer effects, and questionnaire design effects.

## References

Brick, J. and Kalton, G. (1996), "Handling Missing Data in Survey Research," Statistical Methods in Medical Research, 5, 215-238.

Colsher, P.L., and Wallace, R.B. (1989), "Data Quality and Age, Health, and Psychobehavioral Correlates of Item Nonresponse and Inconsistent Response," Journal of Gerontology, Psychological Sciences, 44, 45-52.

Converse, P. (1970), "Attitudes and Nonattitudes: Continuation of a Dialogue," in E.R. Tufte (Ed.), The Quantitative Analysis of Social Problems, Addison Wesley: Reading.

Dillman D. (1978), Mail and Telephone Surveys: The Total Design Method, New York: Wiley.

E. De Leeuw. (1999), "Prevention is the Better Cure: How to Reduce Missing Data," KM, 1999, 20, 39-35.

E. De Leeuw. (1992), "Data Quality in mail, telephone, and face to face surveys," Amsterdam. TT-publikaties.

Gergen, K.J., and Back, K.W. (1966), "Communication in the Interview and the Disengaged Respondent," Public Opinion Quarterly, 30, 385-398.

Groves, R. (1989), Survey Errors and Survey Costs, New York: Wiley.

Herzog, A. and Rodgers, W. (1989), "Age Differences in Memory Performance and Memory Ratings in a Sample Survey," Psychology of Aging.

Lessler, J. and Kalsbeek, W. (1992), Nonsampling Error in Surveys, New York: Wiley.

Lestina F. (2003), "Analysis of the Linguistically Isolated Population in Census 2000," Census 2000 Evaluation A.5.a, Bureau of the Census, Washington D.C.

Raudenbush, S. and Bryk, A. (2002), Hierarchical Linear Models: Applications and Data Analysis Methods, Thousand Oaks: Sage Publications.

Singer J. (1998), "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models," Journal of Educational and Behavioral Statistics, 24:4, 323-355.

Sudman, S., and Bradburn, N. (1974), Response Effects in Surveys, Chicago: Aldine.

Zajac, Kevin. (2003). "Analysis of Imputation Rates for the 100 Percent Person and Housing Unit Data Items from Census 2000." Census 2000 Evaluation B.1.a., Bureau of the Census, Washington D.C.