ANALYSIS OF ITEM ALLOCATION RATES FOR POPULATION ITEMS IN THE CENSUS 2000 DRESS REHEARSAL

Kevin J. Zajac and James B. Treat, U.S. Census Bureau*
Kevin J. Zajac, U.S. Census Bureau, Rm 2501, Bldg 2, Washington DC 20233

Keywords: edit; substitution; imputation

Introduction

The Census 2000 Dress Rehearsal was an operation simulating the processes planned for Census 2000. It took place in three areas of the United States - Sacramento, California; Columbia, South Carolina and surrounding areas; and Menominee, Wisconsin. The operation used questionnaires very similar to those used for Census 2000. Completed questionnaires containing respondent information from the Dress Rehearsal were data captured by computer and the records were placed into an unedited computer file. Some data were later edited and imputed. This paper focuses on item allocation rates of the 100 percent population characteristics: relationship, sex, age, date of birth, Hispanic origin, and race. The purpose is to check the completeness of the data received from each of the Dress Rehearsal test sites.

Background

Two methods were used to deliver questionnaires to housing units for Dress Rehearsal. The first method was mailout/mailback, where forms were delivered by the United States Postal Service. This method was used in the largely urban area of Sacramento, California, and in part of the South Carolina test site. The other methodology was update/leave. It relied on Census Bureau enumerators to update maps and addresses as they delivered questionnaires. Update/leave took place in Menominee, Wisconsin, and in a portion of the South Carolina site. The mailout/mailback and update/leave methods both required respondents to fill out and mail back their questionnaires. A nonresponse followup (NRFU) operation performed at all three test sites gathered information at housing units which did not return the questionnaire by the cutoff date. During NRFU, enumerators filled out the questionnaires by personally visiting the housing units.

The two types of enumeration and NRFU resulted in two questionnaire return types, self-administered and enumerator-administered, during Dress Rehearsal. Each return type had long and short form versions. The self-administered forms were specifically designed to be completed and returned by the respondent. They were used in the mailout/mailback and update/leave areas. Approximately one out of every six housing units received a long form mail return questionnaire. Every housing unit was designated a long or short form prior to the start of the mailout/mailback and update/leave operations. Because these same designations remained for the housing units in the NRFU universe, enumerators had information so that they knew which units should be enumerated with a long form questionnaire. Long form questionnaires contained several additional housing unit and person questions than their short form questionnaire counterparts. However, there were only seven common person questions between these two forms. These questions covered name, relationship, sex, age, date of birth, Hispanic origin, and race.

Imputation, which is a process of filling in missing respondent data or replacing existing respondent data, was used during the Dress Rehearsal. The purpose of imputation is to make sure that every housing unit and person item has a value and to maintain certain consistencies among some of these values. Three different components comprise the imputation process: substitution, edit, and allocation.

- Substitution occurs when a full set of characteristics for a person or housing unit needs to be assigned. This happens because a questionnaire contains no information for the household and/or no information for the people within the household. A nearby housing unit with complete information is selected as a substitute and the responses are used to fill the missing data items. This housing unit is selected using a nearest neighbor hot deck.
- An edit is performed when a response for a data item is either missing or not consistent to other responses, and an item value can be determined based on provided information from that same person. For example, the age item can be edited based on date of birth.
- Allocations, or computer assignments of acceptable codes in place of unacceptable entries or blanks, are needed most often when an entry for a given item is lacking or when the information reported for a person or housing unit on that item is inconsistent with other information for that same person or housing unit. This is done by grabbing a response from another person within the household or from a person in a nearby household. For example, if a person was missing a value for sex and it could not be determined using the first name, then allocation would occur to specify a value for the sex variable.

On the file used for this analysis, substitutions do not appear. This was because the substitution processes occurred after the development of this file. However, these substitutions are included on a later created file. Table A shows the number of substituted people as well as total substitution percentages in each site, broken down by short and long form and by self-administered and enumerator-administered form. In Menominee and South Carolina, the percent of substitutions are 1.23 and 2.48, respectively. On average, this means that a little over one person in Menominee and about 2.5 persons in South Carolina, out of every 100 persons at each site, had to be substituted. In Sacramento, the substitution percentage skyrockets to 9.36. This high rate is due to the fact that sampling was done at this test site for vacant units identified by the Postal Service during the form mailout process and by enumerators during NRFU. The characteristics of persons not included in this sample were classified as substitutions.

Table A. Substituted Persons by Test Site, Return Type, and Form Type.

Test Site	Return Type	Short Form Substitutions	Long Form Substitutions	TOTAL Substitutions	% of Persons Substituted
Menominee, Wisconsin	Self	34	9	43	
	Enumerator	4	5	9	
	TOTAL	38	14	52	1.23
Sacramento, California	Self	2835	528	3363	
	Enumerator	25255	3988	29243	
	TOTAL	28090	4516	32606	9.12
South Carolina and surrounding areas	Self	6797	1156	7953	
	Enumerator	5897	1111	7008	
	TOTAL	12694	2267	14961	2.48

Methodology

The data used for this analysis reflect characteristics of persons living or staying in housing units from the three Dress Rehearsal test sites. Excluded from this data are any persons who were substituted during the imputation process. Altogether the analysis covered 324,801 persons in Sacramento, 589,364 persons in South Carolina, and 4,159 persons in Menominee. The information came from about 150,000 housing units in Sacramento, about 250,000 housing units in South Carolina, and about 1,700 housing units in Menominee.

For the purpose of this paper, the term 'response(s)' will refer to 'non-allocated response(s)'. This embodies any responses that were edited as well as responses that were unchanged by the imputation process. Thus, a value that has not been allocated for an item will have a 'response'.

Several questions will be addressed in this paper regarding the completeness of the data from the Dress Rehearsal. These questions include:

- *What are the overall item allocation rates for each 100 percent population item by test site?*
- *Do the overall item allocation rates differ between the Menominee, Sacramento, and South Carolina test sites?*
- *Are item allocation rates different between long and short form questionnaires for each 100 percent population item by test site?*
- *Are item allocation rates different between self-administered and enumerator-administered questionnaires for each 100 percent population item by test site?*
- *Are item allocation rates different between the combinations of long and short form and self-administered and enumerator-administered questionnaires for each 100 percent population item by test site?*

Item allocation rates indicate the portion of a population in which an allocation occurred for a particular question. In this analysis, item allocation rates are computed for the six 100 percent population items - sex, relationship, age, date of birth, Hispanic origin, and race. These items are dissected between each Dress Rehearsal test site by long and short form, by self-

administered and enumerator-administered form type, and by these two together. The item allocation rate will be computed as follows:

$$R_{ti} = [(N_{ti} - V_{ti}) / N_{ti}] * 100;$$

where R = Item Allocation Rate
N = number of nonsubstituted persons
V = number of nonsubstituted persons with a nonallocated response
t = test site (1=Menominee, 2=Sacramento, 3=South Carolina)
i = item number (1=relationship, 2=sex, 3=age, 4=date of birth, 5=Hispanic origin, 6=race)

Each person level record was classified by form length (long/short) and form return type (self-administered/enumerator-administered) according to the questionnaire which was completed by the respondent.

The wording of the 100 percent population questions between long and short forms was virtually the same. However, between the self-administered and enumerator-administered forms, the age and date of birth items were different. On the enumerator form, there is a note telling the enumerator to ask the age question if the date of birth is not known. This note does not appear on the self-administered forms.

On the file used for this analysis, there were allocation flag variables created for the 100 percent population items to show whether the respondent's answer was a valid response or if it was allocated in any way. If a person's response was allocated according to an allocation flag variable, the item was tagged as having an item allocation. Every variable had different criteria to determine what constituted a response.

- For relationship item, it was considered to be nonallocated when a response category was marked off, a value was reported in a write-in field, or a reported value was changed for household consistency.
- The sex question had a nonallocated response when a category was marked off or when sex was determined by the first name of the person.
- Age had a nonallocated response when the age was reported, when the date of birth was reported, or when age and date of birth were both reported and were corresponding.
- The date of birth item was considered to be valid when there was a valid numerical response in the month, day, and year parts of the write-in field. Additionally, when the year portion of the date and either the month or day portion of date are accepted, it was also considered to be a nonallocated response.
- A response for Hispanic origin was considered to be nonallocated when anywhere from 1 to 3 origin categories were marked for the item, or when it was assigned from the race code.
- The race variable had a nonallocated response only when at least one category was marked or when a valid race was written in the write-in box.

Results

Table 1 shows the overall item allocation rates for the 100 percent population items by Dress Rehearsal test site. Date of birth and Hispanic origin have considerably greater item allocation rates than the other four items in Menominee. Similarly, South Carolina has a high item allocation rate for each of these two items. In Sacramento, date of birth and Hispanic origin, as well as race, produce larger item allocation rates than the other three 100-percent items. The date of birth and race item allocation rates in Sacramento are considerably greater than the other two sites. Meanwhile, the relationship, sex, age, and Hispanic origin items have rates that are comparable across the three test sites. The rates of these four items differ by no more than 1 percentage point between site.

Table 1. Overall Item Allocation Rates for 100 Percent Population Items by Dress Rehearsal Test Site.

	Dress Rehearsal Test Site		
Item	Menominee	Sacramento	South Carolina
Rel.	1.8	2.2	1.9
Sex	0.3	0.2	0.2
Age	1.2	1.8	1.9
DOB	6.4	11.5	9.4
Hisp.	4.9	4.2	5.2
Race	0.6	5.4	1.2

(Item abbreviations: Rel.=Relationship; DOB=Date of Birth; Hisp.=Hispanic Origin)

Table 2 gives the item allocation rates for the 100 percent population items by Dress Rehearsal test site and by form length (long form versus short form). In every case, relationship, age, and date of birth produce greater item allocation rates on long forms within each test site. Hispanic origin and race, in contrast, have higher rates on short forms for all test sites. Of all 100 percent population items, date of birth had the largest item allocation rate at each test site for both short and long forms. The 1.6 percentage point difference between long

and short form item allocation rates in Sacramento on the date of birth item was the largest difference between form lengths of all three sites.

Table 2. Item Allocation Rates for 100 Percent Population Items by Dress Rehearsal Test Site and Form Length.

	Dress Rehearsal Test Site and Form Length					
	Menominee		*Sacramento*		*South Carolina*	
Item	Long	Short	Long	Short	Long	Short
Relationship	2.43	1.72	3.01	2.12	2.55	1.76
Sex	0.41	0.30	0.20	0.26	0.24	0.20
Age	1.42	1.20	2.53	1.74	2.47	1.82
Date of Birth	7.51	6.30	12.92	11.31	9.80	9.37
Hispanic Origin	4.67	4.96	3.54	4.33	4.56	5.32
Race	0.20	0.63	4.33	5.51	1.08	1.16

Table 3 indicates the item allocation rates for the 100 percent population items by Dress Rehearsal test site and by form return type (self-administered versus enumerator-administered). According to this table, the item allocation rates for enumerator-administered forms are noticeably greater than self-administered within all three sites for relationship, age, and especially date of birth. The date of birth item produced the largest item allocation rates on enumerator-administered forms across all sites. In Sacramento, the item allocation rate for date of birth on enumerator-administered forms was nearly 30 percentage points greater than self-administered forms. The difference in item allocation rates between self-administered and enumerator-administered forms for the date of birth item is over 22 percentage points in South Carolina. In Menominee, this difference is over 10 percentage points. The Hispanic origin item, meanwhile, produced higher item allocation rates on self-administered forms across all three sites.

Table 3. Item Allocation Rates for 100 Percent Population Items by Dress Rehearsal Test Site and Form Return Type.

	Dress Rehearsal Test Site and Form Return Type					
	Menominee		*Sacramento*		*South Carolina*	
Item	Self	Enum	Self	Enum	Self	Enum
Relationship	1.41	2.31	1.50	3.97	1.41	2.84
Sex	0.43	0.17	0.23	0.29	0.18	0.26
Age	0.73	1.87	0.83	4.23	0.79	4.23
Date of Birth	1.92	12.27	2.60	32.47	2.23	24.32
Hispanic Origin	8.63	0.17	5.06	2.27	6.81	1.89
Race	0.64	0.50	6.65	2.32	1.15	1.16

Table 4 gives the item allocation rates in Menominee for the 100 percent population items by form return type and form length. For all self-administered forms and most enumerator-administered forms, item allocation rates between long and short forms differ by less than 1 percentage point for each item. The only noticeable exception is the rate between enumerator-administered long and short forms for the relationship item (4.26 to 2.02). Comparing across form return types, long form enumerator-returns had greater item allocation rates than self-administered long forms for relationship, sex, age, and date of birth. As well, enumerator-administered short

forms had a higher item allocation than self-administered short forms for the relationship, age, and date of birth items. In looking only at the date of birth item, the enumerator-administered long and short forms have item allocation rates that are clearly larger than self-administered long and short forms, respectively. This is consistent with what was observed in the overall comparison between self-administered and enumerator-administered forms (Table 3). For the Hispanic origin and race items, item allocation rates on self-administered returns were greater than enumerator-administered returns on both long and short forms. This is also consistent with the results between self-administered and enumerator-administered forms in Table 3.

Table 4. Item Allocation Rates for 100 Percent Population Items by Form Return Type and Form Length in Menominee, Wisconsin.

	Form Type and Form Length			
	Self		*Enumerator*	
Item	Long	Short	Long	Short
Rel.	0.78	1.49	4.26	2.02
Sex	0.00	0.48	0.85	0.06
Age	0.39	0.77	2.55	1.77
DOB	2.33	1.87	13.19	12.14
Hisp.	8.53	8.64	0.43	0.13
Race	0.39	0.67	0.00	0.57

Table 5 points out the item allocation rates in Sacramento for the 100 percent population items by form return type and form length. For relationship, sex, age, and date of birth, the item allocation rate for self-administered returns differs by less than 0.4 percentage points between long and short forms. However, self-administered short forms have item allocation rates that are nearly two percentage points greater than self-administered long forms for the Hispanic origin and race items. On enumerator-administered returns, the item allocation rate on long forms is larger than short forms for every 100 percent item except date of birth. Results from Table 2 indicated that long forms had larger item allocation rates than short forms on the relationship, age, and date of birth items, but not on the sex, Hispanic origin and race items. The greatest difference between form lengths in enumerator-administered forms occurred on the relationship item, where the item allocation rate was 5.35 percent on long forms and about 3.7 percent on short forms. Across form return type, Hispanic origin and race showed that self-administered long and short forms had larger allocation rates than long and short form enumerator-administered returns, respectively. This is similar to Table 3, which breaks down item allocation rates between self and enumerator-administered forms. Meanwhile, there were lower item allocation rates for self-administered long forms than there were for enumerator-administered long forms on the relationship, sex, age, and date of birth items. As well, self-administered short forms had item allocation rates that were lower than, or nearly even with, enumerator-administered short forms for these same population items. These rates are also consistent with the overall comparison between self-administered and enumerator-administered forms in Table 3. Date of birth has an exceptionally large rate of item allocation for both enumerator-administered long and short forms. Approximately one in three long and short form enumerator returns received in Sacramento had an item allocation for the date of birth item.

Table 5. Item Allocation Rates for 100 Percent Population Items by Form Return Type and Form Length in Sacramento, California.

	Form Type and Form Length			
	Self		*Enumerator*	
Item	Long	Short	Long	Short
Rel.	1.79	1.46	5.35	3.72
Sex	0.05	0.26	0.51	0.24
Age	0.97	0.81	5.49	4.00
DOB	2.71	2.59	32.36	32.49
Hisp.	3.67	5.26	3.29	2.08
Race	4.97	6.89	3.10	2.17

Table 6 indicates the item allocation rates in South Carolina for the 100 percent population items by form return type and form length. Except for Hispanic origin, the item allocation rates for self-administered long and short forms are within about 0.5 percentage points. For Hispanic origin, the rates differ by about 1.2 percentage points. On every item except date of birth, the difference in item allocation rates for enumerator-administered long and short forms is less than about 1 percentage point. The item allocation rates of long and short form enumerator-administered returns differ by about 3.4 percentage points for the date of birth item. This is somewhat uniform with the results from the overall comparison between long and short forms in Table 2. In comparing across form return type, self-administered long forms have lower item allocation rates than enumerator-

administered long forms for every item except Hispanic origin. Likewise, self-administered short forms have smaller allocation rates than short form enumerator-returns for relationship, sex, age, and date of birth. Between self and enumerator-administered long forms, nearly a 4 percentage point difference in item allocation rates existed for the age item, about a 1.7 percentage point difference for the relationship item, and about a 3.2 percentage point difference for the Hispanic origin item. Between self and enumerator-administered short forms, there is about a 3.3 percentage point difference in item allocation rates for age and about 1.3 percentage points for the relationship item. For Hispanic origin, the difference was greater than 5.2 percentage points. Additionally, the item allocation rates on enumerator-administered long and short forms for the date of birth item are higher than the respective self-administered forms by around 20 percentage points. These results follow consistently with Table 3 observations comparing item allocation rates from self and enumerator-administered forms.

Table 6. Item Allocation Rates for 100 Percent Population Items by Form Return Type and Form Length in South Carolina.

	Form Type and Form Length			
	Self		*Enumerator*	
Item	Long	Short	Long	Short
Rel.	1.89	1.34	3.61	2.67
Sex	0.08	0.20	0.49	0.21
Age	0.95	0.77	4.92	4.09
DOB	2.54	2.18	21.52	24.90
Hisp.	5.78	6.97	2.59	1.74
Race	0.72	1.22	1.65	1.06

Conclusions

Item allocation rates between self-administered and enumerator-administered forms from Dress Rehearsal were considerably different in most cases. The relationship, age, and date of birth items consistently showed that enumerator returns had higher rates of item allocation within each test site. The date of birth item was the most significant, showing close to a 30 percentage point increase in item allocation rates from self-administered to enumerator-administered forms (see Table 3). This large rate difference may have also been caused by the fact that age and date of birth questions were included under the same person question on the enumerator questionnaire in Dress Rehearsal. This question lists the date of birth item above the age item, but instructs the enumerator to get the person's age if the date of birth is not known or incomplete. Some enumerators may have asked only the age question to speed up the interview since it requires less writing on the form.

Information from the data capture system revealed that enumerators were using pencil to complete the form. It was found in several cases that the marks made by the pencil were not properly data captured. This is a likely cause for the high enumerator item allocation rates.

On the other hand, the Hispanic origin and race items generally indicated a higher rate of item allocation on the self-administered forms than the enumerator-administered forms. Although the self-administered questionnaires specifically make a note to alert the respondent to answer both the Hispanic origin and race questions, there may have been confusion because of the similarity of the items. More analysis may be needed to determine a solid answer for the high item allocation rates on self-administered forms for the Hispanic origin and race items.

Form length had little effect on the analysis. The item allocation rates on the long forms were not substantially different than those produced by short forms. This similarity in item allocation rates may be somewhat expected since the wording of the six 100 percent population items on the long and short forms was identical.

References

Jones, C.D. (1991). "Additional Information on Count and Content Quality in the 1990 Census." Internal Census Memorandum, May 22, 1991.

Bureau of the Census. (1992). Accuracy of the Data section in 1990 Census of Population and Housing: Summary Population and Housing Characteristics. Washington, D.C.: Bureau of the Census.

Acknowledgements

The authors would like to thank the following people for their contributions to this paper: Deborah Griffin of the Decennial Statistical Methods Division, and Michael Ikeda of the Statistical Research Division for their review of the paper.

* 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.