

Introducing Parametric Models and Administrative Records into 2014 SIPP Imputations

Gary Benedetto
Joanna Motro
Martha Stinson

U.S. Census Bureau
4600 Silver Hill Road
Suitland, MD 20746

Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

As part of the redesign of the Survey of Income and Program Participation (SIPP), we have developed a new imputation process for handling respondents who are missing entire sections of data. This new process creates indicator variables called topic flags that determine whether each section of questions was relevant for a respondent (e.g. receipt of Food Stamps). We model the joint distribution of these variables and covariates from the survey and from independent data sources using a parametric method called Sequential Regression Multiple Imputation (SRMI). The resulting, approximate distribution is used to impute missing values for the topic flags. Modeling topic flags in this manner is an alternative imputation method to hot-deck imputation or whole-record donation for cases where respondents did not complete the majority of the survey. As opposed to hot-deck imputation that can only control for a limited number of characteristics, the SRMI approach is able to control for many more variables, including household, parent, and spouse characteristics. Moreover, our process incorporates administrative records for the first time into SIPP production. These data offer a valuable set of covariates whose availability is unrelated to survey non-response and, as a result, helps to mitigate problems caused when survey data are not “missing at random”. This paper describes our modeling process, its advantages over more traditional imputation methods like hot-deck imputation, and demonstrates the usefulness of linking administrative data into the models. Lastly, we show preliminary results for the SIPP 2014 panel topic flags.

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion. The proportions and other statistics in the text, figures, and tables of this report are presented here for the purpose of evaluating new imputation methodology. The weighting and design effects necessary to interpret the reported statistics as estimates of underlying population parameters are not available at this time. Apparent differences may not be statistically significant, but all comparative statements in this report have undergone statistical testing and are significant at the 95% confidence level. All data are subject to error arising from a variety of sources, including sampling error, nonsampling error, model error and any other sources of error. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Introduction

In 2006, the Census Bureau undertook a major redesign of the Survey of Income and Program Participation (SIPP). The redesign included changing the length of the survey reference period, introducing an event history calendar instrument, completely overhauling the data processing system that produces public use data, and using linked administrative records to improve the quality of the data. As part of this redesign effort, new imputation methods were explored that utilized both the linked data and more modern techniques to improve the handling of missing data. This paper explains how the model-based imputation methods that grew out of this research were adopted to impute high-level topic screener indicators in the 2014 SIPP, the first production panel to utilize the redesigned system. We present the results of a select number of imputed topic indicators to show how the new imputation method is an improvement over past methods. Administrative data was also used to correct responses where people confused receipt of Social Security Insurance (SSI) and Old-Age, Survivors, and Disability Insurance (OASDI). We present the basic results of this correction for SSI. A detailed report of the correction can be found in Giefer, Williams, Bendetto, and Motro 2015.

Background and Improvements of New Imputation Method

Previously, SIPP processing handled missing data by using a hot deck to choose a donor. The donor's answers were then used to fill in the data for the respondent with missing values. For respondents missing answers to only a few questions, the hot deck imputation was done on a variable by variable basis. In cases where the respondent failed to answer enough questions for a sufficient partial interview, one donor was selected to provide a full-record donation. A major disadvantage of this hot deck method was only being able to control for limited characteristics when creating donor cells, due to the need to make the cells sufficiently large. This heightened the perennial concern about data not missing at random conditional on the hot deck stratifying variables. In the case of whole record donation, any information about a topic that was provided by the respondent was over-written and the resulting record was sometimes inconsistent with data reported by other household members.

The new imputation process seeks to address these concerns by adopting a model-based approach that is able to control for many more variables, including household, parent, and spouse characteristics and information from administrative records. Topic flags are now modeled the same way for everyone, whether a respondent is missing one topic or all of them. Imputation of downstream variables still relies on hot decks but the universe for these variables is built off the topic flag with model-based imputations.

Topic Flags

A topic is a higher-order level question in a block of SIPP questions about a specific content area. Most often, these questions screen respondents to determine whether they should answer subsequent questions about a particular content area. For example, a screener question might ask if a person received OASDI benefits. If the respondent answers 'yes,' they are asked subsequent questions about their OASDI benefits, including when they started and how much they received. If the respondent answers 'no', they bypass the additional questions and move on to the next content area. If the respondent fails to answer the higher-order/screener question, then a response is imputed through the model-based process. For example, for those respondents who did not answer the labor force section of the survey, we imputed whether they held any jobs at any point during the reference year. Similarly, for those who did not answer the social welfare benefits questions, we imputed section screener questions such as Temporary Assistance for Needy Families (TANF), Supplemental Nutrition Assistance Program (SNAP), and Supplemental Security Income (SSI) receipt. Table 1 shows a full list of topics and percent of data imputed per topic.

Topic flags are created prior to imputation and contain valid 'yes' and 'no' responses or flagged as either 'needs imputing' or 'missing and not in universe'. After the imputation is complete, topic flags have no missing data where a respondent is in universe for a topic. Each topic flag has an accompanied allocation flag indicating if the response comes from a valid response, a model-based imputation, a logical imputation¹, or is not in universe.

¹ Logically imputed responses are those where information was provided in other SIPP questions that inform a 'yes' or 'no' response to a topic flag.

Topic flags serve as an alternative imputation method to hot-deck imputation or whole-record donation for respondents who missed entire sections of the SIPP. The process of imputing topic flags also incorporates administrative data to mitigate problems caused when survey data are not “missing at random.” These topic flags then facilitate downstream question edits, determining who is in universe to answer more detailed questions about the specific content area. For people who are in universe for downstream questions but fail to answer those questions, those responses are imputed using a hot-deck method.

Data

SIPP

The SIPP is a nationally representative household-based survey focused on sources of income, jobs, and government-based assistance programs. The model-based imputation was first implemented for households that were interviewed in 2014 using the re-designed SIPP survey instrument to ask about their economic well-being in 2013. This paper focuses on the first wave of the 2014 SIPP panel.

In order to reduce respondent burden, some topics were asked to only one person in a household or family member within a household on behalf of other household/family members. In this paper, for these topics, we discuss only the imputation for people who were supposed to answer the question on behalf of the household or family clump² within a household, not necessarily the equivalent of people who were covered by certain programs.

Administrative Data

The model based imputation makes use of six different sources of data shared with the Census Bureau by the Social Security Administration (SSA). First, we use two types of earnings records derived from W-2 forms filed with SSA by employers. The Detailed Earnings Record (DER) extract reports uncapped income-taxable earnings for each employer that filed a W-2 record from 1978-2012. It also contains a report of earnings that were not income taxable and were deferred into accounts like 401(k) plans. We utilize the DER to create a measure of total earnings in a given year and to count the number of jobs an individual held. From the DER we also create measures of self-employed earnings and an indicator of any deferred earnings. We also utilize the Summary Earnings Record (SER) extract which contains total earnings capped at the FICA taxable maximum from 1951-2012 to create a count of how many years an individual has worked over his or her lifetime.

Next, we make use of the Master Beneficiary Record (MBR) and Payment History Update System (PHUS) extracts to create indicators for whether an individual was eligible for and received OASDI payments due to retirement, disability, spouse retirement or death, parent retirement or death, or some combination of reasons. These extracts contain both present benefit receipt and historical information so we are able to tell what year an individual started receiving benefits and whether they ever stopped. The Supplemental Security Record (SSR) provides the same information about SSI benefits. These three files combined together give us a very accurate picture of who was receiving OASDI benefits, SSI benefits, or both. This information in turn is very helpful in predicting reports of OASDI and SSI receipt.

Finally, we make use of the Numident, a register of all Social Security Numbers (SSNs) ever issued in the United States, along with the MBR and SSR, as an administrative source of birth date information. If a person is receiving benefits we utilize the birth date from the benefits files in order to create an age for the individual during the survey reference period. If the person is not receiving benefits, we use the birth date from the Numident. While this does not replace the survey reported age on the final public use data, we do use this age derived from administrative data as an explanatory variable in our models.

Methodology

Sequential Regression Multivariate Imputation

² Age and immediate family members of the household respondent define a clump within a household. For example, if a father is the household respondent, his spouse, children under 22, and him are in a clump. For some SIPP questions, the father (household respondent) answers on behalf of his spouse and children under 22. In this scenario, grandparents and any children 22+ years old living in the household are not in the clump and respond to questions themselves.

Once the topic flags have been created and the observations with no survey response have been flagged for imputation, we are faced with the challenge of estimating a joint probability distribution of the full set of topic flags conditional on all of the survey and administrative data we have available to us. Specifically, suppose that X is the collection of explanatory variables and Y_1, Y_2, \dots, Y_k are the k topic flags all of which contain some with missing values. We need to estimate $p(Y_1, Y_2, \dots, Y_k | X)$. If the missing data pattern were monotonic (i.e., the topic flags could be arranged such that Y_n is missing whenever Y_m is missing for all $n > m$) then we could simply decompose the joint distribution into a sequence of conditional marginal distributions,

$p(Y_1, Y_2, \dots, Y_k | X) = p(Y_1 | X) p(Y_2 | Y_1, X) \dots p(Y_{k-1} | Y_1, \dots, Y_{k-1}, X)$. However, since missing data patterns in surveys tend to be non-monotonic, we need a more sophisticated approach to make use of all the non-missing data in our model.

The sequential regression multivariate imputation (SRMI) method offers an intuitive, relatively easy to implement, and computationally low cost approach to estimating the joint distribution described above (Raghunathan et al., 2001). The SRMI uses an iterative approach with a sequence of regressions to allow all of the non-missing data to eventually ‘soak’ into the model. In the first iteration, we use the sequence of conditional marginals mentioned above to initialize the SRMI with completed data. In other words, we regress Y_1 on X , and use the results of that regression to make an initial imputation of the missing values in Y_1 .³ We call this initial, completed Y_1 vector, $Y_1^{(1)}$. Then we regress Y_2 on X and $Y_1^{(1)}$ and impute a value for $Y_2^{(1)}$. We continue doing this until the first iteration is complete, ending with the data vector $(Y_1^{(1)}, Y_2^{(1)}, \dots, Y_k^{(1)})$.

As mentioned earlier, because of the non-monotone missing data pattern, this initial set of completed values fails to condition on all of the non-missing data. We continue to iterate through the list regressing Y_m on $(X, Y_1^{(t)}, \dots, Y_{m-1}^{(t)}, Y_{m+1}^{(t-1)}, \dots, Y_k^{(t-1)})$ for $m = 1, \dots, k$ and $t = 2, \dots, T$ where T is the last iteration.⁴ Since the regressions in these iterations condition on the entire array of the most recently completed data, the non-missing data that went unused in iteration 1 increasingly influences the full joint distribution of variables. While there is no specific convergence criterion, empirical analysis has shown that fewer than 20 and generally as few as 5 to 10 iterations are sufficient to condition the imputed values in any variable on all other variables (Ambler and Royston, 2007; van Buuren, 2007; He et al., 2009).

SSI and OASDI Program Confusion Correction

An early review of the Social Security Insurance (SSI) topic for reported data showed potential program confusion with OASDI. Both SSI and OASDI provide benefits to people who are blind, disabled or over a specified age. However, people must meet different requirements to receive either or both programs. Linking respondents to administrative data showed that almost one-half of respondents who reported receiving SSI did not have corresponding administrative data showing SSI receipt. However, many of these people did have administrative data showing OASDI receipt. Therefore, the decision was made to implement a program confusion correction after the imputation process. The correction can change both valid and imputed responses. The basics of the correction are:

- 1) In cases where the SSI topic flag indicated program receipt but the administrative data indicated *only* OASDI receipt, we change the SSI topic flag to ‘no.’ One exception to this fix is when a respondent reports receiving *state* SSI benefits. When there is reported SIPP data indicating receipt of state SSI benefits we keep the ‘yes’ response to the SSI topic flag because we only have federal and federally administered administrative SSI records, not administrative data on state SSI receipt (which can be different from federal SSI).
- 2) In addition to the SSI topic flag correction, we also make an OASDI topic flag correction when someone reports receiving SSI but administrative records only indicate OASDI receipt. The concern here is that someone might have reported getting SSI and then did not report receiving OASDI because they thought they had provided the information on OASDI already. In cases where someone reports receiving SSI but the administrative data shows only OASDI receipt, we correct the OASDI topic flag to match the administrative records.

³ Since the topic flags are binary variables, we use a logistic regression.

⁴ For this paper, $T = 5$.

- 3) The same process is employed in the reverse for OASDI benefits. Individuals who reported OASDI benefit receipt but had no administrative record of receipt are compared to SSI administrative data. If they in fact received SSI benefits, the OASDI topic flag is set to 'no' and the SSI topic flag is set to 'yes.'

Results

Table 1 shows basic summary statistics of the percent of missing data needing imputation, by topic imputed through SRMI, and the person-level match rate to administrative records. For most topics, about 3% to 7% of respondents in universe for a topic are missing and are flagged for imputation. For each topic, about 65% to 70% of respondents flagged to be imputed match to administrative records. This high match rate suggests that administrative records are likely a good source of independent data to use in the model-based imputation. The final column of Table 1 shows the percent of people in universe for each topic who did respond per topic and matched to administrative records. Here we see that about 90% of in-universe respondents for each topic matched to administrative records. It is not surprising that people who responded to the survey were also more likely to be matched to administrative records because the probabilistic matching is likely higher for responders (e.g., better quality matching variables, people who do not move often, etc.).

After the five iterations of SRMI and the post-imputation fix to address SSI-OASDI program confusion, we performed some data review to test the quality of the imputations and analyze the value of the administrative data in our models. Table 2a shows demographic characteristics for Job Line 1. In the re-designed SIPP, respondents are asked about jobs they held in the previous year. In this paper, we focus on the first job reported on Job Line 1. Table 2b shows demographic characteristics for receiving SNAP benefits. SNAP benefits are only asked of clump respondents (defined by family relationships and age) and people 15 years and older who were not in clumps. The percentages in these tables are column percentages. The first two columns of each table show the percent of people in universe for the topic who responded to the topic and those who were missing and flagged to be imputed. The last two columns are people who responded 'yes' to having a Job in 2013 (Table 2a) and 'yes' to receiving SNAP benefits in 2013 (Table 2b).

One of the most obvious problems with past imputations in the SIPP has been the inconsistency between earnings and income-qualified program participation. This has largely been due to hot-deck procedures not being able to condition on a very large set of covariates. While the new imputation methods have only been implemented for high-level topic flags, we have reason to believe that these models, which allow many more covariates to influence the imputations, will help address this problem.

The SNAP topic flag is a good example of maintaining consistency across earnings and income-qualified program participation with the new imputation method. In Table 2c, we see SNAP participation broken down by category of total 2012 household administrative earnings. The first two columns show the percent of people who reported 'yes' received SNAP benefits and those imputed to have received SNAP benefits in 2013. For households that have administrative record earnings of \$25,000 or more, there is seemingly a big difference in 38% of imputed households receiving SNAP while only 22% of non-imputed households reported receiving SNAP. Upon further inspection, this large percentage point difference is possibly due to differences in the earnings distribution between responders and non-responders for everyone in-universe for the SNAP question (the second two columns). In these second two columns, a larger percentage of non-responders were in the \$25,000 or more category compared to the percentage of responders (68% and 55%, respectively). In the final two columns of Table 2c that conditions on being in-universe for the SNAP topic, 5.3% reported 'yes' received SNAP benefits and 6.7% were imputed to 'yes' received SNAP benefits. Therefore, the differences in the distribution of household earnings between imputed SNAP participants and reported SNAP participants (first two columns) are due mainly to observable differences in the distribution of household earnings for non-responders and responders.

The difference in administrative data between the missing and non-missing populations is a good example of how having an independent source of data on the survey respondents helps to address the classic problem of survey data not missing at random. Perhaps the clearest example of this is with the employment indicators. Table 3 shows a cross-tabulation of the presence of a job in the first job line of the survey with the indicator for positive earnings in the 2012 administrative data. We see that 59.6% of job line 1 responders say they have a job in the survey and 58% have positive earnings in the 2012 administrative data. For the non-responders, we impute 62.9% to have a job in job line 1, which is 3 percentage points higher than the rate for responders. However, the non-responders also have

positive earnings in the 2012 administrative data at a rate of 61.7% which is also about 3 percentage points higher than what we see for the responders. Without this extra source of covariation in our model, it is not clear that the imputed rate for the presence of a job would be different from the responders despite the real presence of differences (observed through the administrative records, but unobserved in the survey data) between the two populations.

As mentioned in the Methodology section, we were also able to use the administrative records in a combined edit-imputation to correct for obvious program confusion. After the SRMI but prior to the program confusion edit, Table 4 shows that 1559 respondents have positive participation in SSI in the survey but do not have SSI receipt in the administrative data implying a false positive rate of 48.2%. The post-imputation edit reduces this cell dramatically to 329 resulting in a much more acceptable false positive rate of 15.6%.

Finally, while there is a dearth of diagnostics for the quality of imputations, Raghunathan and Bondarenko (2007) suggest a promising statistic that is fairly intuitive. They suggest using the full, completed data to estimate a logistic regression for whether a surveyed individual responds to an item or not. If a strong model can be estimated to predict response, then one can compare the imputed and non-imputed values of the variable in question for respondents with similar propensities to respond. If the imputations are good under the missing-at-random assumption (or missing-at-random conditional on a good set of independent covariates such as the administrative records we use), then the distributions of imputed and reported values for individuals with similar response propensities should also be similar. Table 5 shows the rate of the “Yes” response for 5 topic flags broken down by imputed versus reported for each quintile of the predicted propensity of response for that topic. The five topics (Job Line 1, SNAP, WIC, Disability, and Education Enrollment) were chosen because of sufficient sample sizes and because we were able to estimate response models which did not fail the Hosmer-Lemeshow goodness-of-fit test.⁵ For most cells, the imputed and reported means are not statistically different. For all but one of the cells where the means are statistically different, the magnitude of the difference is small from an analyst perspective. For example, for Job Line 1, quintile 2, roughly 72% of the responders have a job and roughly 68% of non-responders are imputed to have a job. That difference, while statistically significant, is quite small in magnitude. The one major problem cell is the 2nd quintile of the propensity to respond for SNAP. This result indicates that the SNAP imputation model may need further refinement in future iterations of modeling.

Conclusion

The introduction of parametric models conditioning on survey and administrative data into the missing data imputation process is a great step forward for the SIPP. Problems with not missing-at-random have been addressed. A much larger set of covariates, from both the survey and the administrative records, can now influence the imputations of very important, high-level screeners. Linking to administrative data also allows for corrections in survey data when there is strong evidence of program confusion. We hope that the improvements made at this high level will also benefit the more traditional edits and imputations for the more detailed questions downstream of the topic flags.

In addition to benefiting the SIPP, these imputation methods have the potential to benefit other Census products as well. This work shows that it is possible to obtain and integrate administrative records into survey processing systems in a timely manner and that these outside data sources are highly valuable in identifying differences between individuals that are not apparent from the survey data. Other surveys could implement similar methods and reduce the bias that comes from the violation of the missing-at-random assumption.

⁵ The Hosmer-Lemeshow test is a goodness-of-fit test for logistic regressions. It has some limitations in that rejecting the null implies a bad model fit, but not rejecting the null does not necessarily mean it is a good model. Moreover, if the sample is too small, we may not reject the null even with a poor model.

References

- Ambler G, Omar RZ, Royston P. 2007. "A comparison of imputation techniques for handling missing data predictor values in a risk model with a binary outcome." *Statistical Methods in Medical Research* 16:277–298.
- van Buuren S. 2007. "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical Methods in Medical Research* 16: 219–242.
- Giefer, Katherine, Williams, Abby, Benedetto, Gary, and Motro, Joanna, 2015. "Program Confusion in the 2014 SIPP: Using Administrative Records to Correct False Positive SSI Reports." Forthcoming in *FCSM 2015 Proceedings*.
- He, Yulei, et al. 2009. "Multiple imputation in a large-scale complex survey: a practical guide." *Statistical Methods in Medical Research* 19 (6): 653-670.
- Raghunathan, Trivellore, Lepkowski, James M., van Hoewyk, John, and Solenberger, Peter. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology* 27 (1): 85-96.
- Raghunathan, Trivellore and Bondarenko, Irina, 2007. "Diagnostics for Multiple Imputations." *SSRN Working Paper Series*. Available at SSRN: <http://ssrn.com/abstract=1031750> or <http://dx.doi.org/10.2139/ssrn.1031750>

Table 1: Summary of Topic Flag Data

Topic	Total in Universe	Percent of in- Universe Respondents Imputed	Percent of in- Universe, <i>Imputed</i> Respondents who Linked to AdRecs	Percent of in- Universe, <i>Reported</i> Data Respondents who Linked to AdRecs
Disability	58090	6.0	68.1	91.7
Disability Payments	9946	34.2	68.0	94.3
Disability Limitations	72401	7.5	70.7	91.2
Energy Assistance	18614	4.5	81.1	92.1
Fertility	58367	4.3	66.4	91.3
Lump Sum Payments	58030	6.1	68.1	91.7
Retired	43857	4.8	69.3	92.9
Retirement Payments	13048	16.2	68.6	96.0
Spousal Payments	4607	3.6	76.3	95.3
OASDI-Self	55148	5.8	67.3	91.7
OASDI-Kids	20454	8.7	70.5	89.6
Unemployment Compensation	58090	6.0	68.1	91.6
Veterans Benefits	4999	3.6	73.2	95.9
Worker's Compensation	58050	6.0	69.0	91.6
Education Enrollment	70441	6.0	69.3	91.2
SNAP	41011	5.0	61.8	91.0
General Assistance	41011	4.9	61.6	91.0
Job 1	58367	5.6	68.1	91.5
Job 2	58367	5.8	69.3	91.5
Private Health Insurance	73215	6.4	65.7	91.0
Military Health Insurance	73215	7.5	67.1	91.1
Medicare Health Insurance	73215	7.7	67.4	91.1
Medicaid Health Insurance	73215	10.2	73.6	91.1
SSI* (pre-correction)	72065	6.3	69.5	91.2
Dependent Care	20835	2.1	72.0	92.9
TANF	41011	4.9	61.7	91.0
WIC	58367	3.4	70.5	90.9

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.

Table2a: Demographic characteristics of respondents in universe for the Job 1 screener questions; respondents who answered having a job and imputed to have a job in 2013

	All Respondents		Respondents with a Job in 2013	
	SIPP Job 1 Reported	SIPP Job 1 Missing	SIPP Job 1 Reported	SIPP Job 1 Imputed
	Col. %	Col. %	Col. %	Col. %
Sex				
Male	47.3	51.4	51.5	54.0
Female	52.7	48.6	48.5	46.0
Age				
15-25	15.9	31.9	12.8	24.7
26-49	38.5	40.6	51.3	51.6
50-64	25.6	19.3	29.0	20.9
65+	20.0	8.2	7.0	2.8
Race				
White	65.1	55.3	66.0	57.7
Black	13.4	15.7	12.4	14.2
Hispanic	6.8	10.2	6.6	9.5
Other	14.7	18.8	15.1	18.7
Marital Status				
Married	48.3	35.4	52.8	39.7
Not Married	51.7	64.6	47.2	60.3
Education				
No HS	18.4	22.7	10.4	13.6
HS	29.4	31.7	29.1	31.9
SoCo	19.6	18.9	21.0	20.9
Assoc	8.1	6.4	9.7	8.0
College	15.6	14.2	19.5	17.9
Grad	9.0	6.2	11.3	7.7
Household Income from Admin Records*				
No Earnings	18.2	6.7	3.5	1.8
Under \$25,000	22.7	21.4	21.4	18.7
\$25,000 and Above	59.1	71.9	75.1	79.6
N	55170	3197	32865	2011

*Administrative earnings lagged one year (2012) because of availability. Imputed where missing.

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.

Table 2b: Demographic characteristics of respondents who answered the SNAP screener questions; respondents who answered receiving SNAP benefits and imputed as receiving SNAP benefits in 2013

	All Respondents		Respondents Receiving SNAP in 2013	
	SIPP SNAP Reported	SIPP SNAP Missing	SIPP SNAP Reported	SIPP SNAP Imputed
	Col. %	Col. %	Col. %	Col. %
Sex				
Male	46.6	52.9	29.0	38.6
Female	53.4	47.1	71.0	61.4
Age				
15-25	12.0	29.3	10.8	29.6
26-49	40.8	45.7	51.2	43.2
50-64	25.9	16.9	23.8	20.0
65+	21.3	8.1	14.3	7.3
Race				
White	63.8	51.0	45.4	39.6
Black	15.0	17.0	27.9	25.9
Hispanic	6.7	10.9	6.6	7.7
Other	14.6	21.1	20.0	26.8
Marital Status				
Married	37.2	17.2	18.5	12.7
Not Married	62.8	82.8	81.5	87.3
Education				
No HS	15.3	19.9	30.5	28.6
HS	30.4	34.5	37.9	43.2
SoCo	20.9	19.6	20.3	16.4
Assoc	8.4	6.4	5.7	4.6
College	16.0	14.6	4.2	X
Grad	9.0	4.9	1.4	X
Household Income from Admin Records*				
No Earnings	20.5	6.7	35.3	16.4
Under \$25,000	24.9	25.0	43.0	45.9
\$25,000 and Above	54.6	68.4	21.7	37.7
N	39003	2008	5300	220

*Administrative earnings lagged one year (2012) because of availability. Imputed where missing.

"X" indicates suppressed for disclosure reasons.

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.

Table 2c: 2013 SNAP Benefit Topic Flag Response and Receipt by 2012 Earnings from Administrative Records

	“Yes” Received SNAP in 2013		In-universe for SNAP		Conditional “Yes” Percentages on In-universe for SNAP	
	SIPP Reported	SIPP Imputed	SIPP Reported	SIPP Missing	SIPP Reported	SIPP Imputed
	Col %	Col %	Col %	Col %		
Household Income from Administrative Records						
No Earnings	35.3	16.4	20.5	6.7	23.5	22.4
<\$25,000	43.0	45.9	24.9	25	23.5	19.8
>=\$25,000	21.7	37.7	54.6	68.4	5.3	6.7
Total	100	100	100	100		

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.

Table 3: Job Line 1 Topic Flag compared to 2012 AdRec Employment

Overall Percentages for cases where SIPP respondent answered the first question about jobs held (94.5% of in-universe respondents)			
Percent worked for pay in 2013?		Percent with W-2/Schedule C positive earnings in 2012?	
Yes	59.6	Yes	58.0
No	40.4	No	42.0

Overall Percentages for cases where SIPP respondent DID NOT answer the first question about jobs held and TF was imputed (5.5% of in-universe respondents)			
Percent worked for pay in 2013?		Percent with W-2/Schedule C positive earnings in 2012?	
Yes	62.9	Yes	61.7
No	37.1	No	38.4

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Detailed Earnings Record (DER), Summary Earnings Record (SER), and Numident.

Table 4a: Distribution of Pre- and Post-Correction SIPP SSI Receipt by Administrative Record SSI Receipt: 2013. (only respondents successfully linked to Administrative Records)

SIPP SSI Topic Flag	Administrative Record SSI Receipt: "Yes"	Administrative Record SSI Receipt: "No"
	N	N
Before Correction		
Yes	1674	1559
No	535	61603
After Correction		
Yes	1778	329
No	431	62833

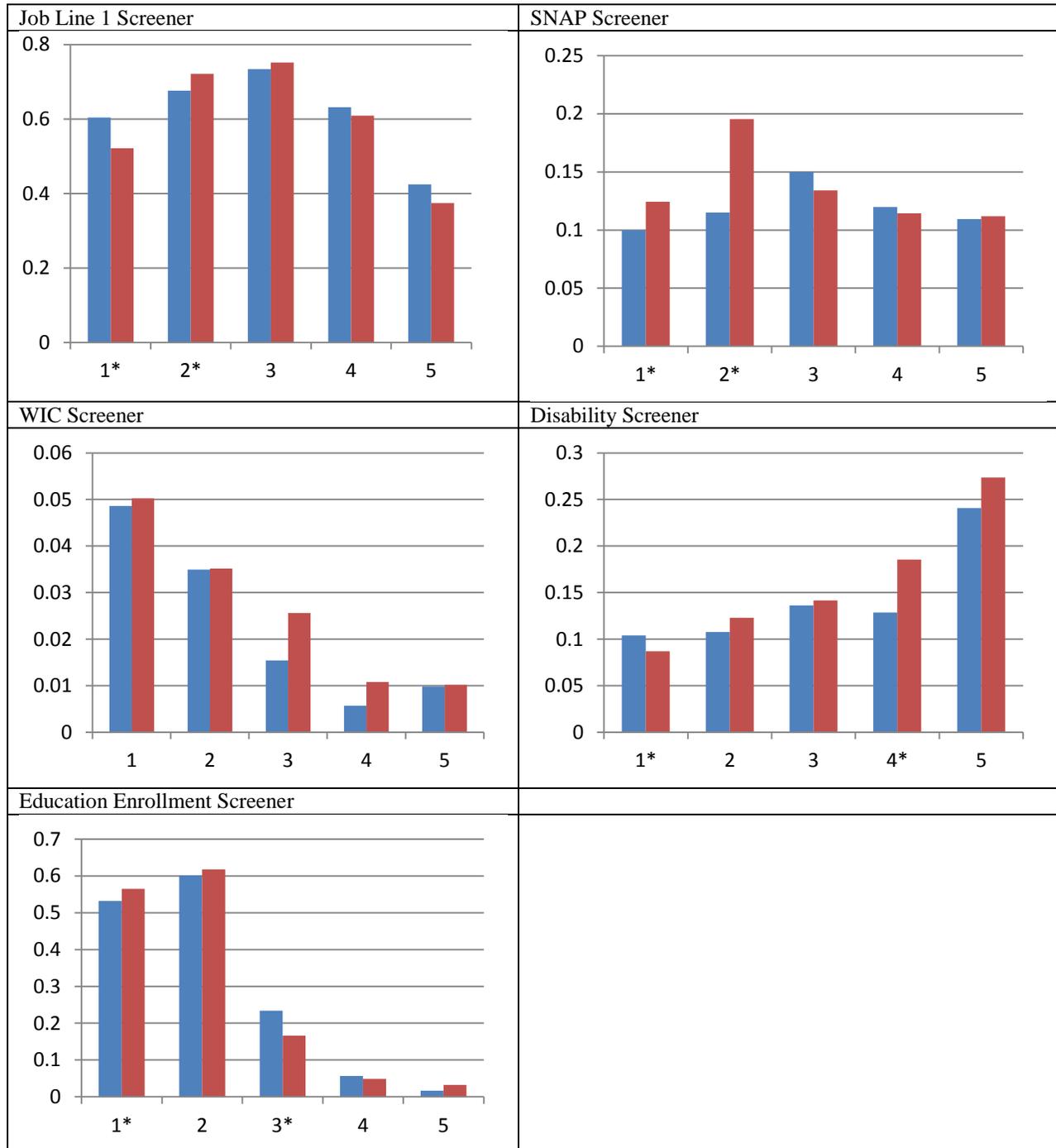
Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.

Table 4b: Distribution of Pre- and Post-Correction SIPP OASDI Receipt by Administrative Record SSI Receipt: 2013. (only respondents successfully linked to Administrative Records)

SIPP OASDI Topic Flag	Administrative Record OASDI Receipt: "Yes"	Administrative Record OASDI Receipt: "No"
	N	N
Before Correction		
Yes	11329	821
No	2157	51064
After Correction		
Yes	11668	524
No	1818	51361

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.

Table 5: Rate of “Yes” by Quintile of Predicted Response Propensity. In each pair, left (blue) bars represent imputed cases for non-responders, and right (red) bars represent actual responses.



*indicates the difference in mean between imputed and non-imputed cases is statistically different at 95% confidence level.

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel (Wave 1) and administrative records from the Master Beneficiary Record (MBR), Supplemental Security Record (SSR), Detailed Earnings Record (DER), Payment History (PHUS), Summary Earnings Record (SER), and Numident.