Imputation of Multiple-Response Items in SESTAT and Its Component Surveys

Nicholas Beyler[1], Bruno Vizcarra[1], Donsig Jang[1]
Amang Sukasih[1]
[1]Mathematica Policy Research, 600 Maryland Ave., SW, Washington, DC 20024

Abstract
The Scientists and Engineers Statistical Data System (SESTAT) is a comprehensive, integrated system of information about the characteristics of scientists and engineers in the United States. Unweighted sequential hot-deck imputation is used to handle item nonresponse in each of the SESTAT component surveys. To maintain consistency across surveys and survey years, stepwise regression models are fit to determine control variables and the control variables are used to identify donors for hot-deck imputation. Many items in the SESTAT component surveys contain multiple, potentially-correlated questions, yet the current imputation protocol calls for the responses to these questions to be imputed separately, without accounting for the correlation structure in the outcomes. To evaluate the validity of this approach, we consider an alternative method for imputing multi-response items, where control variables are determined using the Sequential Regression Multivariate Imputation (SRMI) method, which sequentially fits regression models to account for correlation among the outcome variables. Imputed responses based on the current protocol are compared to those based on this alternative approach using data from one of the SESTAT component surveys. Preliminary findings show differences between the distributions of imputed responses when comparing the two approaches.

Key Words: Hot-Deck Imputation, Multivariate Imputation, SESTAT, *IVEware*

1. Introduction

In the survey setting, nonresponse comes in the form of unit nonresponse and item nonresponse. Unit nonresponse occurs when a sampled unit does not complete the entire questionnaire because of refusal, inability to complete the questionnaire, or other factors. Typically, sampling weights are adjusted to account for unit nonresponse using information from the sampling frame (Groves et al. 2002; Lessler and Kalsbeek 1992). Item nonresponse occurs when a sampled unit completes the questionnaire but leaves one or more questions unanswered because the question is too sensitive, the answer to the question is unknown, the question is skipped by mistake, or other factors. Typically, missing responses to questions due to item nonresponse are imputed using information from other answered questions (Kalton and Kasprzyk 1986).

One of the most frequently used imputation methods is hot-deck imputation, whereby a missing response to a question from one sampled unit (the recipient) is replaced with the observed response to the question from another sampled unit (the donor). For the purposes of imputation, all respondents (recipients and donors) are often partitioned into imputation classes based on variables that are observed for all respondents and are also highly correlated with the variable being considered for imputation. Then, respondents within these classes are often sorted based on additional variables highly correlated with the variable being imputed. The donor for a recipient is chosen as the respondent with an observed value for the imputed variable who is also nearest to the recipient after sorting (e.g., the nearest neighbor). See Andridge and Little (2010) for a more comprehensive introduction to hot-deck imputation.

In practice, there are often multiple survey variables that have missing values and the missing values follow a general pattern across respondents. Judkins (1997) calls this situation a "swiss cheese pattern" for missing data. Performing hot-deck imputation under this situation is often difficult when the variables with missing values are correlated with one another. Simply imputing the variables one at a time in some pre-specified order or imputing the variables together in a single imputation pass may not adequately account for the correlations between the variables and could lead to biased survey estimates based on imputed data.

In this paper, we address the problem of performing hot-deck imputation when there is a general pattern of missing data. Using data from the 2006 National Survey of Recent College Graduates (NSRCG) – one of the SESTAT component surveys – we perform hot-deck imputation on multiple-response items (i.e., items with related, potentially-correlated responses) in two different ways. First, we impute the multiple variables individually in a pre-specified order as is currently done in practice. Second, we impute the multiple variables in a three-step approach that accounts for the correlation among the multiple responses. In Section 2, we provide some general information on the current hot-deck imputation procedures for SESTAT and its component surveys. In Section 3, we describe our proposed three-step approach for performing hot-deck imputation, which makes use of the Sequential Regression Multivariate Imputation (SRMI) method developed by Raghunathan et al. (2001). In Section 4, we present some preliminary imputation results based on 2006 NSRCG data for two multiple-response item examples. We conclude with a summary of the results in Section 5.

2. Imputation for SESTAT and Its Component Surveys

2.1 Background

The Scientists and Engineers Statistical Data System (SESTAT) is a comprehensive, integrated data system sponsored by the National Science Foundation (NSF) that provides users with employment, educational, and demographic information about scientists and engineers in the United States. SESTAT is comprised of data from the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR). Each of these component surveys are conducted every few years and the most recent survey years are 2006, 2008, and 2010. Each of the SESTAT component surveys collect similar information from potentially overlapping segments of the population composed of all individuals in the United States that are 75 years or younger as of the survey reference date and hold a bachelor's or higher degree in science, engineering, or a related health field. NSCG covers most of this population, while NSRCG focuses on the most recent college graduates with a bachelor's or master's degree in science or engineering and SDR focuses on college graduates with a doctorate degree in science or engineering. Visit the

SESTAT website for more information on these component surveys and survey eligibility (www.sestat.nsf.gov).

2.2 SESTAT Hot-Deck Imputation

Beginning in 2003, NSF has developed imputation specifications (or guidelines) for all three component surveys so that missing values due to item nonresponse are imputed consistently across the surveys before the data are incorporating into SESTAT. All SESTAT items subject to item nonresponse are imputed using an unweighted sequential hot-deck imputation procedure (Beyler et al. 2011; Jang and Lin 2009). As part of this procedure, *control variables* are determined for each item subject to imputation. The control variables consist of *classing variables* and *sorting variables*.

The classing variables define classes (or cells) within which all imputation must occur. In most cases, classing variables are *filter variables*, which determine whether a subsequent question or set of questions in the survey should be asked for a particular respondent. For example, the filter variable employment status (e.g., are you currently working, yes or no?) is used as a classing variable for imputing salary so that individuals who are not currently working are not subject to imputation of salary.

The sorting variables are used to sort cases within the imputation classes prior to imputation. Sorting variables are determined using stepwise multivariate regression analysis. Variables subject to imputation are regressed on potential sorting variables and the most significant variables from the final fitted stepwise regression models are used as sorting variables for imputation. Potential sorting variables must be fully observed for all individuals and are often stratification variables used for the SESTAT surveys, such as citizenship status, degree year, field of major, race/ethnicity, and gender. Variables that have already been imputed may also be considered as sorting variables for variables that have yet to be imputed. For example, some of the potential sorting variables for imputing salary are gender, age, race/ethnicity, job field, and year of degree. A stepwise regression model regressing salary on these potential sorting variables is fit and age and job field may be two significant variables in the model based on some pre-specified criteria (e.g., the model coefficient p-value is less than 0.20). These two variables – age and job field – would then be used to sort cases within imputation classes before imputing salary.

The survey contractors perform hot-deck imputation using in-house programs that utilize the control variables listed in the SESTAT imputation specifications (Beyer et al. 2011; Jang and Lin 2009). For items with multiple sorting variables, serpentine sorting is implemented so that adjacent records are similar with respect to as many sorting variables as possible (Beyler et al. 2011; Jang and Lin 2009). After sorting is complete, a missing value for the item subject to imputation is replaced with a value from the last encountered donor in the list. For the most part, the items subject to imputation are imputed in the order in which they appear in the survey (some demographic and degree-related items are imputed first so they can be used as sorting variables for imputing other items).

2.3 Hot-Deck Imputation for Multiple-Response Items

Many items in the SESTAT component surveys have multiple-response options. For example, one question asks “Did these factors influence your decision to work in an area outside the field of your highest degree?” and then lists 7 different factors that respondents must mark as either “yes” or “no” (Figure 1). Another question asks “For which of the following reasons did you take training during the past 12 months?” and then lists 7 different reasons that respondents must mark as either "yes" or "no" (Figure 2). We refer to these kinds of items as *multiple-response items*.

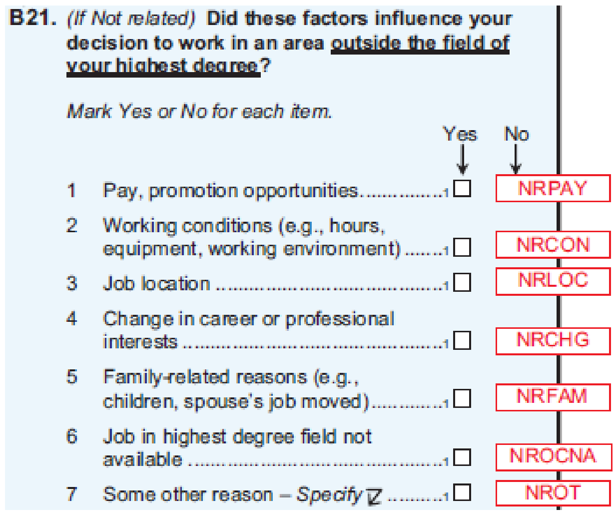

Figure 1: Multiple-response item from a SESTAT component survey asking about factors that influenced the decision to work in an area outside the field of highest degree

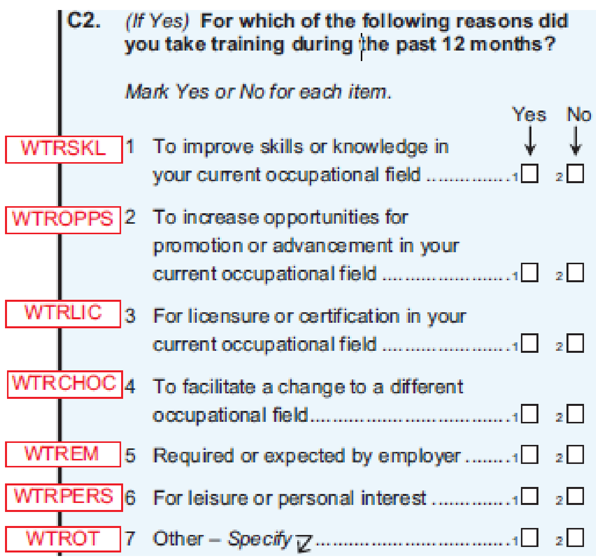

Figure 2: Multiple-response item from a SESTAT component survey asking about reasons for taking training during the past 12 months

The SESTAT imputation specifications provide the survey contractors with two general options for imputing multiple-response items. For multiple-response items with 4 or more response options like the ones shown in Figures 1 and 2, contractors are instructed to impute the responses separately, where multiple donors may be used to replace multiple missing values for a single recipient. For example, if a recipient is missing two responses, one of the responses may be imputed from donor A and the other response may be imputed from donor B using two different hot-deck imputation passes with two different sets of control variables. For other multiple-response items with 3 or less response options, contractors are instructed to impute the response items together so that a single donor is always used to replace the multiple missing values for a single recipient. In this paper, we focus on the first option which is applicable for multiple-response items with 4 or more response options.

One concern with imputing the responses from a multiple-response item individually using multiple hot-deck imputations with multiple sets of control variables is that responses that are imputed in a later pass cannot be used as control variables for the responses that are imputed in an earlier pass, even if the responses are highly correlated with one another. For example, suppose that the variables listed in Figure 1 are imputed in seven passes where NRPAY is imputed first, NRCON is imputed second, and so on. Under this scenario, even if the variable NRPAY is highly correlated with the variable NRCON, the variable NRCON cannot be used as a control variable for imputing NRPAY because it is imputed after NRPAY.

Ideally, the hot-deck imputation process should not limit the choice of control variables that can be used for imputation so that the correlations between the variables from a multiple-response item are fully accounted for during imputation. An alternative hot-deck procedure should allow for the scenario where NRCON is used as a control variable when imputing NRPAY even if NRCON is imputed after NRPAY. Developing such an approach is difficult under the hot-deck imputation framework and may require multiple hot-deck iterations (Andridge and Little 2010).

3. Sequential Regression Multivariate Imputation (SRMI)

3.1 Standard SRMI Imputation

In order to develop an alternative hot-deck imputation approach that more fully accounts for the correlations among the variables from a multiple-response item we consider the regression imputation method proposed by Raghunathan et al. (2001) called Sequential Regression Multivariate Imputation (SRMI). See Kalton and Kasprzyk (1986) for a general introduction to regression imputation. To illustrate the SRMI method for a sample of size n, let $Y_1,\ldots,Y_k$ be a set of k variables with a general pattern (or "swiss cheese pattern") of missing values and assume, without loss of generality, that the variables are ordered so that Y_1 has the smallest amount of missing values and Y_k has the largest amount of missing values. Let X be an n x p dimensional design matrix with no missing values which may include an intercept and other variables that are thought to be correlated with the set of Y variables. The joint conditional density of $Y_1,\ldots,Y_k$ may be written as

$$f(Y_1,Y_2,\ldots,Y_k|X,\theta_1,\theta_2,\ldots,\theta_k) = f_1(Y_1|X,\theta_1)f_2(Y_2|X,Y_1,\theta_2)\cdots f_k(Y_k|X,Y_1,\ldots,Y_{k-1},\theta_k),$$

where f_j, is the conditional density function for Y_j and θ_j is a vector of model parameters (e.g., regression coefficients) in the conditional distribution f_j for $j = 1,\ldots,k$.

The imputation for $Y_1,\ldots,Y_k$ is performed in rounds. Round 1 starts by regressing Y_1 on X and imputing the missing values for Y_1 using the fitted regression model. Next, Y_2 is regressed on (X,Y_1) where Y_1 is the "imputed version" of Y_1 and missing values for Y_2 are imputed. This process is repeated until the last step of round 1 where Y_k is regressed on $(X,Y_1,\ldots,Y_{k-1})$ and imputed using the fitted model. In round 2, the same general process is implemented except that Y_1 is regressed on $(X,Y_2,\ldots,Y_k)$ where $Y_2,\ldots,Y_k$ are the "imputed versions" of $Y_2,\ldots,Y_k$ from round 1, Y_2 is regressed on $(X,Y_1,Y_3,\ldots,Y_k)$ where Y_1 is the "imputed version" of Y_1 from round 2 and $Y_3,\ldots,Y_k$ are the "imputed versions" of $Y_3,\ldots,Y_k$

from round 1, and so on. The imputation concludes at the end of a pre-specified number of rounds or after there is reasonable stability in the values being imputed.

Under this process, missing values for $Y_1,\ldots,Y_k$ are imputed by taking into account (1) the correlation between the Y variables and a set of observed covariates (X) and (2) the correlations between the actual Y variables themselves. The SRMI method is implemented using *IVEware* statistical software (Raghunathan et al. 2002) which is capable of performing single or multiple imputations, and imputations for continuous, binary, count, or a mixed set of variables.

3.2 SRMI for Hot-Deck Imputation

The SRMI method described in Section 3.1 is a regression imputation method, not a hot-deck imputation method. However, the SRMI methodology may be used for the purpose of performing hot-deck imputation. In this section, we propose a method which incorporates the SRMI methodology into the current SESTAT hot-deck imputation framework.

For a sample of size n, let $Y_1,\ldots,Y_k$ be a multiple-response item with k response options (i.e., k variables) that are subject to a general pattern of missing data and let X be an n x p dimensional design matrix for $Y_1,\ldots,Y_k$ where each row of X is the same and represents a potential set of sorting variables that may be used for imputing the Y variables. A three-step approach is considered for imputing $Y_1,\ldots,Y_k$ using hot-deck imputation.

In step 1, the variables $Y_1,\ldots,Y_k$ are imputed using the standard SRMI approach implemented with *IVEware*. Using information from the *IVEware* output, the variables that are most significant in the final round of imputation for each of the Y variables are identified as the sorting variables for each of those Y variables. For example, the most significant variables use for imputing Y_1 (including any X variables and other Y variables) are identified, the most significant variables used for imputed Y_2 (including any X variables and other Y variables) are identified, and so on. This first step is not used to determine the final imputations, but instead is used to determine a set of sorting variables that adequately accounts for the correlations among the Y variables.

In step 2, the variables $Y_1,\ldots,Y_k$ are imputed in an initial hot-deck imputation pass. Without loss of generality, let Y_1 be the first variable imputed, let Y_2 be the second variable imputed, and so on. Each variable is imputed using only the sorting variables determined during step 1 that have no missing values prior to imputation. For example, if Y_2 was determined as a sorting variable for Y_1 during step 1, it may not used for imputing Y_1 in step 2 because Y_2 will have missing values prior to imputation of Y_1 in step 2. The purpose of step 2 is to provide a preliminary set of nonmissing Y variables which are based only on a subset of the sorting variables determined during step 1.

In step 3, the variables $Y_1,\ldots,Y_k$ are imputed in a final hot-deck imputation pass. The variables are imputed in the same order as step 2 (Y_1 first, Y_2 second, etc.) but are now imputed using the full set of sorting variables determined during step 1. In step 2, Y_2 could not be used as a sorting variable for imputing Y_1. But in step 3, the version of Y_2 imputed during step 2 can be used as a sorting variable for imputing Y_1. Step 3 uses the full set of sorting variables determined during step 1 to perform hot-deck imputation on the set of variables $Y_1,\ldots,Y_k$ and thereby accounts for the correlations among the Y variables during imputation.

4. Exploratory Analyses

In this section, we present some preliminary results that compare the imputation results based on the standard hot-deck imputation approach currently implemented for the SESTAT component surveys and the proposed hot-deck imputation approach outlined in Section 3.2 that uses SRMI to produce sorting variables for hot-deck imputation. The goal of these exploratory analyses is to determine (1) if there are differences in terms of the distributions of imputed values using these two hot-deck imputation approaches, and if so, (2) what factors are causing these differences. In this section, we also present the imputation results based on the standard SRMI regression imputation method, but we focus on the differences between the standard hot-deck approach and the modified hot-deck approach.

Imputation for the two multiple-response items that are shown in Figures 1 and 2 are considered using data from the 2006 NSRCG. For the first example presented in Section 4.1, there were a total of 2,809 cases asked to respond to this multiple-response item and 75 of those cases had missing responses that required imputation. For the second example presented in Section 4.2, there were a total of 10,185 cases asked to respond to this multiple-response item and 266 of those cases had missing responses that required imputation. Due to an editing step prior to imputation, every case with missing responses in both examples had missing responses for all the variables from the multiple-response item. In both examples, the percent of cases with missing responses is less than 3 percent, and because of this low item nonresponse, we will compare the results based on the imputed cases only, and not the full set of responses.

4.1 Example 1

First, we consider imputation for the multiple-response item displayed in Figure 1, which asks respondents to respond “yes” or “no” to 7 questions about the factors that influenced their decision to work in an area outside the field of their highest degree. The missing responses to these questions were imputed using the existing (standard) hot-deck imputation approach, the standard SRMI regression imputation approach, and the modified hot-deck imputation approach described in Section 3.2. A single, pre-specified set of potential sorting variables was considered for each imputation method so that the differences in imputations across methods can be attributable to the actual methods and not the choices of sorting variables.

The imputation results are given in Table 1. For each of the variables (NRPAY, NRCON,…,NROT), we give the percent of imputed cases that had an imputed response of “yes” depending on which imputation method was used. For example, 54.7% of the missing values for NRPAY were imputed as yes under the standard hot-deck imputation approach, 64.0% were imputed as yes under the standard SRMI approach, and 53.3% were imputed as yes under the modified hot-deck approach. There are clearly differences in the percentages across the imputation methods verifying that the choice of imputation method will lead to different imputations for this multiple-response item.

The differences between the standard hot-deck approach and the modified hot-deck approach are based on the differences in sorting variables that are used for imputation. The sorting variables for the modified hot-deck approach are determined using SRMI and all of the multiple-response items are eligible candidates for sorting variables. On the other hand, the sorting variables for the standard hot-deck approach are determined by fitting separate stepwise regression models for each variable and only multiple-response items that are imputed before the variable subject to imputation are eligible candidates for sorting variables. For example, NRCON is used as one of the sorting variables for imputing NRPAY under the modified hot-deck approach because of the high correlation between these two variables. However, NRCON is not used as a sorting variable for imputing NRPAY under the standard hot-deck approach due to the limitations of this approach – NRCON is imputed after NRPAY and is therefore not an eligible sorting variable candidate for NRPAY.

The differences between the SRMI approach and the hot-deck approaches may be attributable to the fact that SRMI is a regression imputation approach which randomly imputes responses from a model. The differences in imputation rates between the SRMI and hot-deck imputation approaches are not the focus of this paper, but may be a topic of consideration in future work.

Table 1: Percent of Imputed Cases with an Imputed Response of "Yes" for the Multiple-Response Item in Figure 1 ($n = 75$)

Variable	Standard Hot-Deck	Standard SRMI	Modified Hot-Deck
NRPAY	54.7	64.0	53.3
NRCON	53.3	53.3	61.3
NRLOC	49.3	46.7	46.7
NRCHG	28.0	33.3	30.7
NRFAM	12.0	20.0	8.0
NROCNA	46.7	38.7	44.0
NROT	13.3	9.3	9.3

4.2 Example 2

In a second example, we consider imputation for the multiple-response item displayed in Figure 2, which asks respondents to respond "yes" or "no" to 7 questions about the reasons they took training during the past twelve months. As in the previous example, the missing responses to these questions were imputed using the standard hot-deck imputation approach, the SRMI approach, and the modified hot-deck imputation approach. A single set of pre-specified control variables were used for all three approaches.

The percent of imputed cases with an imputed response of "yes" are given in Table 2 for each of the responses based on the three imputation methods. The results are similar to those from Table 1 in that there are clearly differences in the imputation rates for all of the variables across imputation methods, and the differences between the two hot-deck imputation methods are again attributable to the fact that different sorting variables were considered for the modified hot-deck approach and the standard hot-deck approach.

Table 2: Percent of Imputed Cases with an Imputed Response of "Yes" for the Multiple-Response Item in Figure 2 ($n = 266$)

Variable	Standard Hot-Deck	Standard SRMI	Modified Hot-Deck
WTRSKL	94.0	96.2	94.7
WROPPS	55.6	55.6	56.0
WTRLIC	42.1	40.2	39.5
WTRCHOC	10.9	13.5	14.3
WTREM	70.7	65.0	70.3
WTRPERS	43.2	49.2	42.5
WTROT	0.8	0.8	2.3

5. Summary

The existing hot-deck imputation approach used for imputing multiple-response items from the SESTAT component surveys may not fully account for the correlations among the multiple responses. In this paper, we propose an alternative three-step hot-deck imputation approach which more fully accounts for the correlations among the multiple responses during imputation. This modified approach makes use of the SRMI method (Raghunathan et al. 2001) which can be implemented using *IVEware* statistical software (Raghunathan et al. 2002).

There are two main points to take away from the preliminary analyses presented in Section 4 of this paper. The first point is that the results in Tables 1 and 2 clearly indicate that the two competing hot-deck imputation methods lead to different imputation distributions for the multiple-response items that are considered. The second point is that these differences suggest that the correlations among the multiple responses in these examples are not adequately accounted for under the standard hot-deck approach because a different set of control variables is being used for the modified approach which does account for the correlations among the multiple responses using SRMI.

The results presented in this paper are very preliminary in nature, and pave the way for more in-depth analyses. In future work, we will look more closely at whether the modified hot-deck approach is actually superior to the standard hot-deck approach in terms of providing more precision and/or accuracy in the survey estimates which are based on imputed data. If this line of research suggests that the modified approach is superior to the standard approach, the modified approach should be considered for the SESTAT imputation specifications.

Acknowledgements

We would like to thank NSF for providing us with the data necessary to conduct our analyses for this paper.

References

Andridge, R.R., and Little, R.J.A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, 78(1), 40-64.

Beyler, N., Vizcarra, B., Jang, D., and Lin, X. (2011). Imputation Specifications for 2010 SESTAT. Washington, DC: Mathematica Policy Research.

Groves, R.M., Dillman, D., Eltinge, J., and Little, R. (2002). *Survey Nonresponse*. New York: Wiley.

Jang, D, and Lin, X. (2009). Imputation Guidelines for 2008 SESTAT. Washington, DC: Mathematica Policy Research.

Judkins, D.R. (1997). Imputing for Swiss Cheese Patterns of Missing Data. In *Proceedings of Statistics Canada Symposium 1997*.

Kalton, G., and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12(1), 1-16.

Lessler, J.T., and Kalsbeek, W.D. *Nonsampling Errors in Surveys*. New York: Wiley.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27(1), 85-95.

Raghunathan, T.E., Solenberger, P.W., and Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software: User Guide*. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.