12-2007

New Technique for Imputing Missing Item Responses for an Ordinal Variable: Using Tennessee Youth Risk Behavior Survey as an Example.

Andaleeb Abrar Ahmed
East Tennessee State University

New Technique for Imputing Missing Item Responses for an Ordinal Variable: Using Tennessee Youth Risk Behavior Survey as an Example

---

A thesis

presented to

the faculty of the Department of Public Health

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Public Health in Biostatistics

---

by

Andaleeb A. Ahmed

December 2007

Dr. John Wanzer Drane, Committee Chair

Dr. James Anderson, Committee Member

Dr. Timothy Aldrich, Committee Member

Keywords: Missing Data, Ordinal Variable, Imputation

ABSTRACT

New Technique for Imputing Missing Item Responses for an Ordinal Variable: Using Tennessee Youth Risk Behavior Survey as an Example

by

Andaleeb A. Ahmed

Surveys ordinarily ask questions in an ordinal scale and often result in missing data. We suggest a regression based technique for imputing missing ordinal data. Multilevel cumulative logit model was used with an assumption that observed responses of certain key variables can serve as covariate in predicting missing item responses of an ordinal variable. Individual predicted probabilities at each response level were obtained. Average individual predicted probabilities for each response level were used to randomly impute the missing responses using a uniform distribution. Finally, likelihood ratio chi square statistics was used to compare the imputed and observed distributions. Two other forms of multiple imputation algorithms were performed for comparison. Performance of our imputation technique was comparable to other 2 established algorithms. Our method being simpler does not involve any complex algorithms and with further research can potentially be used as an imputation technique for missing ordinal variables.

DEDICATION

I would like to dedicate this thesis to my family and friends who have always been supportive and encouraging.

ACKNOWLEDGEMENTS

My graduate committee chair, Dr. John. W. Drane, has been a great inspiration for me. Your support, guidance, and advice played a pivotal role in the completion of this thesis. I would like to thank you for instilling in me a lifelong passion for statistical sciences.

I would also like to express my gratitude and acknowledge the support and guidance given by Dr. Anderson. As my graduate committee member, your opinion, feedback and comments were extremely crucial in improving, optimizing, and culminating this dissertation process.

Dr. Aldrich, I appreciate the guidance, facilitation, and encouragement you provided during my graduate years at East Tennessee State University. You are not just an exceptional teacher but also a great friend.

CONTENTS

Page

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Incomplete data are a frustrating problem encountered in observational, experimental, and survey research. The best way to handle this problem is to avoid it, but no matter how carefully we collect the data, missing data almost always exist. The two major problems with missing data are: 1) Biased estimates, 2) Reduction of statistical power (inefficient estimates). This problem is even more serious in survey data where unit and item nonresponse are frequent. Ordinal variables are commonly used in sample surveys and often results in significant item non-response. Especially when we work with Human risk behavior surveys (e.g. YRBS, BRFSS, etc.) involving sensitive and stigmatizing questions, the extent of non-response can be significant.

There is often a need to apply a simple and convenient missing data technique (MDT) for these types of ordinal variables. In the last 2 decades there has been a lot of research on handling missing data. Many new techniques for handling missing data have been suggested. The simplest, oldest, and the most intuitive of them, complete case analysis (CC), uses only complete cases for analysis. Traditional ad hoc ways of handling missing data are: list-wise deletion, pair-wise deletion, means substitution, hot deck imputations, and many others. It has been shown in the literature that these techniques do not perform adequately (Graham et al.,1994; Little et al., 1987; and their references).

Some of the newer techniques like multiple imputation (MI), maximum likelihood (ML), and weighting approaches are gaining popularity in recent years. Multiple imputation technique (Rubin, 1987), particularly, is gaining popularity in the statistical fraternity. Books by Little and Rubin (1987), Schafer (1997), and Allison (2002) prove to

be useful resources on multiple imputation and general issues related to missing data. The newer techniques mentioned above require certain assumptions like multivariate normal distribution etc. Departure from these restrictive assumptions can threaten the validity of estimated parameters. Because a dataset can have many types of variables (e.g. Nominal, Ordinal, continuous, etc.), uniformly applying a multivariate normal theory (MVN) is inappropriate. A number of MI based MDTs are available for continuous variables (e.g. regression based techniques, predictive mean matching, propensity score, etc.) but for ordinal variables the choices are limited. Two commonly available MI based MDTs for discrete variables are: Logistic regression method, and Discriminant function method. Both require a *monotone* missing pattern for application. The distribution of ordinal variables can be highly skewed (especially in Human risk behavior studies) and, hence, analyzing them under MVN assumption is not a good idea (Chen et al., 2005).

CHAPTER 2

STUDY PURPOSE

The purpose of this study is to suggest a simple and convenient regression-based single imputation technique for imputing missing ordinal data. Our technique is different from other log-odds based or proportional-odds based imputation technique because it does not involve extensive computing and complicated algorithms. It's easy to implement under a simple assumption. We limit our study to those variables that are either sensitive or stigmatizing for the study population, but it can potentially be extended to any type of ordinal variable. Multilevel cumulative logit model was used with an assumption that observed responses of an ordinal variable can serve as covariate in predicting the missing item responses of an ordinal variable of similar nature. The detailed methodology and application are covered in section 5. Section 3 explains the types of missing data mechanisms, while section 4 gives a brief review of some of the commonly available techniques for handling missing data.

CHAPTER 3

TYPES OF MISSING DATA MECHANISMS

Many of the new and advanced missing data techniques require certain assumptions to be met before they can be used. These assumptions require a thorough understanding of the mechanisms behind "missingness". Based on the work of Rubin (1976, 1987), there are three mechanisms by which the data can be missing: 1) MAR (*missing at random*), 2) MCAR (*missing completely at random*) and, 3) MNAR (*missing not at random* or *nonignorable*). In MAR mechanism, the data points are missing due to another observed characteristic or variable, in other words the missing observation is conditional on some other observed variable. For example, suppose we intend to evaluate the attitude towards a certain health behavior (Q100) and it is known that gender (Q2) affects its response. In this case the missing observations for Q100 are conditional on Q2; hence, it's a case of MAR. MCAR is a special case of MAR where missing observations are not related to any variable. For example, a computer malfunctions and certain observations of a dataset get deleted. In this case the missing observations depend neither on their own value nor on some other observed variable. Complete case analysis in a true MCAR setting will yield unbiased estimates. Third, and the most problematic type of missing data mechanism, is MNAR (missing not at random) or nonignorable nonresponse. Here the missing observations are a function of its own values. In other words the missingness is conditional on its own values. For instance, abused women may be less likely to answer the question on domestic violence and, hence, the missing observations here are dependent on the question itself. None of the available missing data techniques perform well in case of MNAR. Generally variables involving embarrassing,

stigmatizing, and sensitive questions can be assumed to be either missing not at random (MNAR) or missing at random (MAR). While MCAR and MAR are generally *ignorable*, data that is MNAR is *nonignorable* (Rubin, 1976). In case of MNAR, the nonresponders differ systematically from responders and, hence, it can lead to serious bias. Any given data set can have more than one pattern of missing observations.

Another important distinction to be made while handling missing data in a multivariate setting is the pattern of missing data. When there is a stepwise increase in missingness, it is known as *monotone missing pattern.* Consider variables Y_1, Y_2, ..., Y_n (in that order) is said to have a *monotone missing pattern* when the event that a variable Y_i is missing for a particular individual implies that all subsequent variables Yj, $j > i$, are missing for that individual. This pattern of missingness is very uncommon in real life datasets. On the other hand, when partially observed data are *nonmonotone,* the model for missing data points for one variable may take into account the missingness of other covariates. An *arbitrary* pattern of missingness can be either *monotone* or *nonmonotone.*

CHAPTER 4

MISSING DATA TECHNIQUES

Broadly the available missing data techniques fall under one of the five categories with some overlap: deletion, Maximum Likelihood approach, Bayesian approach, weighting methods, and imputation techniques (single & multiple).

Deletion

Listwise and pairwise deletion techniques come under this category. Listwise deletion (complete case analysis) involves removal of cases with partially observed data-points. Pairwise deletion (available case analysis) is similar to listwise deletion except that it uses all available data for pairwise correlational analysis. It has been repeatedly shown in the literature that these techniques generally lead to invalid results. The main problem with deletion techniques are: inflation of type II error (reduced statistical power), and biased estimates. Unless the mechanism of missingness is MCAR, these ad hoc techniques lead to inefficient and biased results. Another ad hoc method for dealing with partially observed covariate data is to drop variables based on their degree of missingness. However, this method might result in dropping some important explanatory variables and hence can lead to model misspecification (Ibrahim et al., 2005).

Maximum Likelihood Approaches

Likelihood based methods like EM algorithm, (Dempster et al., 1977) structural equations, mixed models, etc. have recently been suggested by statisticians. The assumption behind maximum likelihood (ML) based methods is MAR and it is ordinarily

used in logistic and linear regression models. Likelihood function is summarized by averaging a predictive distribution of the missing values. Observed data are used to estimate the parameters and then the estimated parameters are used to estimate the missing values. It assumes that the observed data are a sample drawn from a multivariate normal distribution (MVN).

As mentioned above, this assumption can sometimes be very restrictive. Various computational methods like EM (expectation-maximization) algorithms (Dempster et al., 1977) are needed to maximize the complex likelihood function. Parameters are estimated using full information ML (FIML or simply ML) from available data. Standard errors are obtained from an observed or an expected information matrix. For computational details of ML approach to missing data please refer to Little and Rubin, 1987 and Schafer, 1997. For an in-depth description of EM algorithm and its application please refer to books by Little and Rubin (1987), Shaefer (1997), and McLachlan and Krishnan (1996).

Bayesian Approach

Last decade has seen an increasing use of Bayesian statistics. Basic tenet of Bayesian analysis is the establishment of a prior distribution of probabilities for the estimation of parameters. Here the missing data points are considered as additional parameters to be estimated under the selected prior distribution of the specified model. Multiple imputation can in fact be considered as an alternative expression of Bayesian analysis. Under noninformative prior distribution, the MI and ML approaches closely approximate. Shafer has described these conditions in his book "*Analysis of incomplete multivariate data*".

Weighting Methods

This method of handling missing data is useful in case of Unit nonresponse. The observed cases or units are weighted according to their similarity to nonresponders. It assumes the absence of unit nonresponse bias, but this assumption requires some additional information about nonresponders. These methods lead to decreased sample variance but the standard error calculations become difficult (Little & Rubin, 1989).

Imputation

Imputation is a general term used for "filling-in" or replacement of missing data with plausible values. Imputation can be either single or multiple. When the missing values for a variable are replaced by a single value, it is called single imputation. Mean substitution, hot deck imputation, and last observation carried forward, etc. are types of single imputation. Mean substitution for a variable involves replacement of missing values with the average of its observed values. Hot deck imputation (Ford, 1983; Rizvi, 1983) is a form of single imputation where the missing values for a particular case or respondent is replaced with a value from a similar case or respondent. The US Census Bureau uses this method for its recent population survey. Although hot-deck imputation replaces missing values with realistic values, there is little theoretical reasoning behind its validity.

Important drawbacks of these single imputation techniques are the underestimation of variance and standard error and the assumption of no difference between respondents and nonrespondents. Multiple imputation (MI) (Rubin, 1978, 1987), on the other hand, is a framework under which multiple sets of plausible values are

imputed for a given set of missing values. Posterior predictive distribution is used to repeatedly draw plausible values and *m* completed datasets are created. These *m* multiply imputed, complete datasets are individually analyzed and the results from each analysis are pooled to compute the final parameter estimates (taking into consideration the within-imputation and between-imputation variances). Because both sampling and imputation uncertainties are incorporated in the pooled analysis, the estimates obtained have better theoretical basis. However, the validity of MI depends on method used for generating *m* datasets. MI based techniques requires that the mechanism of missing data is ignorable or MAR (Little et al., 1987).

As mentioned above this assumption is more or less untestable but more variables in the imputation model can make this assumption more plausible (Schafer 1997; Van Buuren el al., 1999,). According to Rubin's recommendation, if imputations are performed under a Bayesian framework, the results of MI can be inferred as approximately Bayesian. For *monotone* missing pattern, both parametric (based on continuous MVN assumption) and nonparametric methods can be applied. For *arbitrary* missing data patterns, the options are quite restrictive. The Markov Chain Monte Carlo method (Shafer, 1997) is a well known parametric method for arbitrary missing patterns. It can be used to impute all missing values or just enough missing values to transform it to a *monotone* missing pattern. Unless there is a large amount of missingness, 3 to 5 multiply imputed datasets are sufficient (Rubin, 1987, p 114).

There are various algorithms of creating multiple imputed datasets like joint modeling or the Imputation-Posterior approach (IP) (Shafer, 1997), expectation maximization importance sampling (EMis) (King el al., 2001), bootstrapping based EM

algorithm (EMB) (Honaker et al., 2006) and fully conditional specification (FCS) (Van Buuren et al. 2006) etc. As of now there is no general consensus on an ideal algorithm.

CHAPTER 5

METHODOLOGY

Data

The data used for this exploratory exercise are from 2005 Tennessee Youth Risk Behavior Survey (TYRBS). Youth Risk Behavior Survey (YRBS) is part of the epidemiologic surveillance system developed by the Center for Disease Control and Prevention to monitor the prevalence of youth behavior that result in the most significant effects on health and well being of youths in United States. Six categories of youth risk behavior are focused in YRBS. These are the behaviors that results in: unintentional and intentional injuries; tobacco use; alcohol and other drug use; sexual behaviors that result in HIV infection, other sexually-transmitted diseases (STDs), and unintended pregnancies; dietary behaviors; and physical activity (Kolbe, 1990).

The Tennessee State Department of Education conducts TYRBS during odd numbered years. YRBS uses a multi stage cluster sample design. The 2005 TYRBS was completed by 1540 students in 45 public high schools in Tennessee during the spring of 2005 (Tennessee Department of Education). The school response rate was 83%, the student response rate was 85%, and the overall response rate was 71%. Students completed a self-administered, anonymous, 87-item questionnaire. Survey procedures were designed to protect the privacy of students by allowing for anonymous and voluntary participation. Local parental permission procedures were followed before survey administration. This survey is weighted and the results can be generalized to all students in Tennessee public schools in grade 9 – 12. However, this is a proposal on statistical methodology and, hence, it is not to be used for inferences where

Epidemiology or Health Behavior is concerned. Table 1 shows the pattern of missing data in our dataset.

Table 1.

Pattern of missing data. "1" means that the variable is observed in the corresponding group and a "0" means that the variable is missing. The table clearly indicates arbitrary missing data pattern.

	Q44	**Q12**	**Q30**	**Q59**	**Q39**	
1256	*1*	*1*	*1*	*1*	*1*	**0**
26	*1*	*0*	*1*	*1*	*1*	**1**
39	*1*	*1*	*0*	*1*	*1*	**1**
138	*1*	*1*	*1*	*1*	*0*	**1**
8	*0*	*1*	*1*	*1*	*1*	**1**
42	*1*	*1*	*1*	*0*	*1*	**1**
2	*1*	*0*	*0*	*1*	*1*	**2**
4	*1*	*1*	*0*	*1*	*0*	**2**
4	*0*	*1*	*1*	*1*	*0*	**2**
1	*1*	*0*	*1*	*0*	*1*	**2**
8	*1*	*1*	*0*	*0*	*1*	**2**
5	*1*	*1*	*1*	*0*	*0*	**2**
1	*0*	*1*	*1*	*0*	*1*	**2**
1	*1*	*0*	*0*	*0*	*1*	**3**
1	*1*	*1*	*0*	*0*	*0*	**3**
1	*0*	*1*	*1*	*0*	*0*	**3**
1	*0*	*0*	*0*	*0*	*1*	**4**
2	*0*	*1*	*0*	*0*	*0*	**4**
Total	**17**	**31**	**58**	**63**	**155**	**324**

Variables

The main goal of our study was to impute the missing values of a key ordinal variable. Five variables were chosen from the 2005 TYRBS dataset, which includes the key ordinal variable (Q39), and four ordinal covariates (Q44, Q59, Q12, and Q30). For the sake of simplicity the sampling design was not taken into account; however, ignoring the sampling design can lead to biased estimates. In practice the readers are advised to incorporate the sampling design using SAS-callable Survey Data Analysis (SUDAAN)

(Shah, Barnwell, & Bieler, 1997), or the latest SURVEYFREQ, SURVEYLOGISTIC, and other survey procedures introduced in SAS/STAT 9.1®. The choices of variables were based on one factor: sensitive and stigmatizing nature of question. Figures 1, 2, 3, 4, and 5 show each variable. Q39 corresponds to the question "*During your life, on how many days have you had at least one drink of alcohol?*" It has 7 ordinal levels corresponding to: 0 days, 1 or 2 days, 3 to 9 days, 10 to 19 days, 20 to 39 days, 40 to 99 days, and 100 or more days.

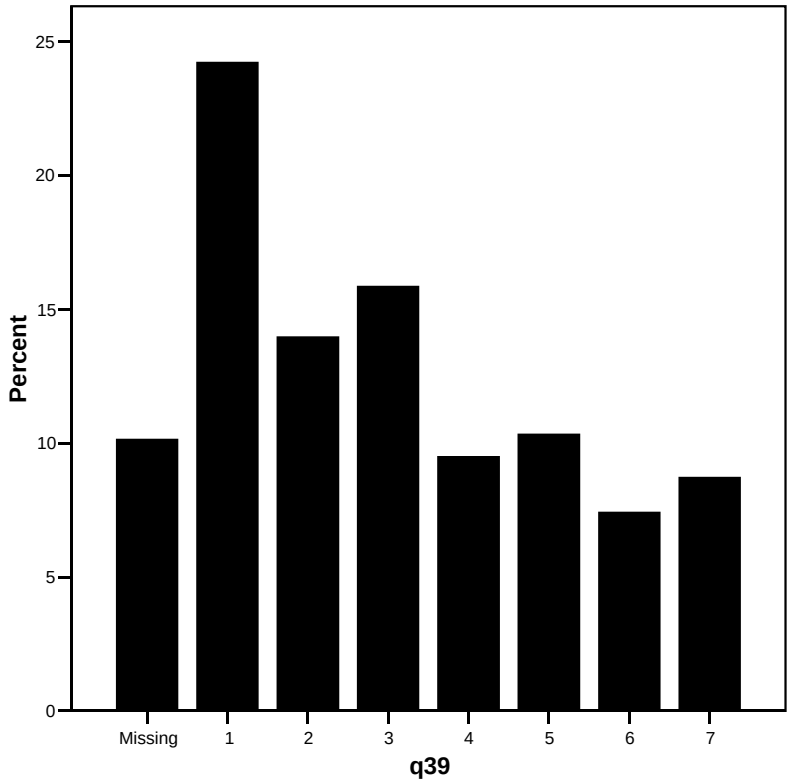

Figure 1. Distribution of response variable (Q39).

Q44 corresponds to the question "*During your life, how many times have you used marijuana?*" It has 7 ordinal levels corresponding to: 0 times, 1 or 2 times, 3 to 9 times, 10 to 19 times, 20 to 39 times, 40 to 99 times, and 100 or more times.

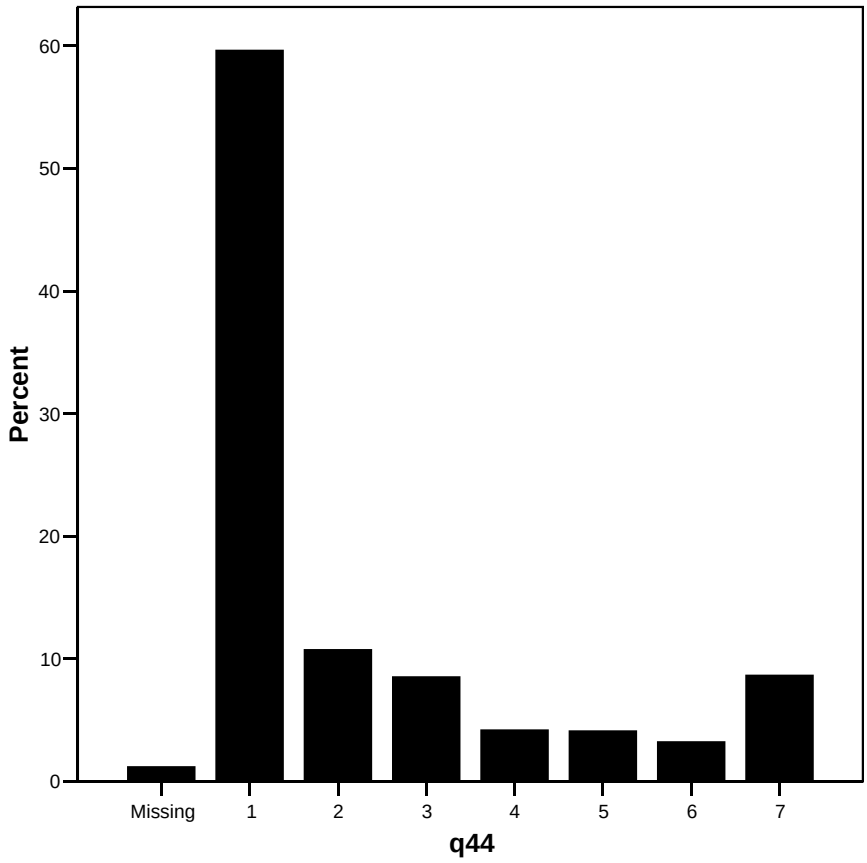

Figure 2. Distribution of response variable (Q44).

Q59 corresponds to the question "*During your life, with how many people have you had sexual intercourse?*" It has 7 ordinal levels corresponding to: Never had sex, 1 person, 2 people, 3 people, 4 people, 5 people, and 6 or more people.

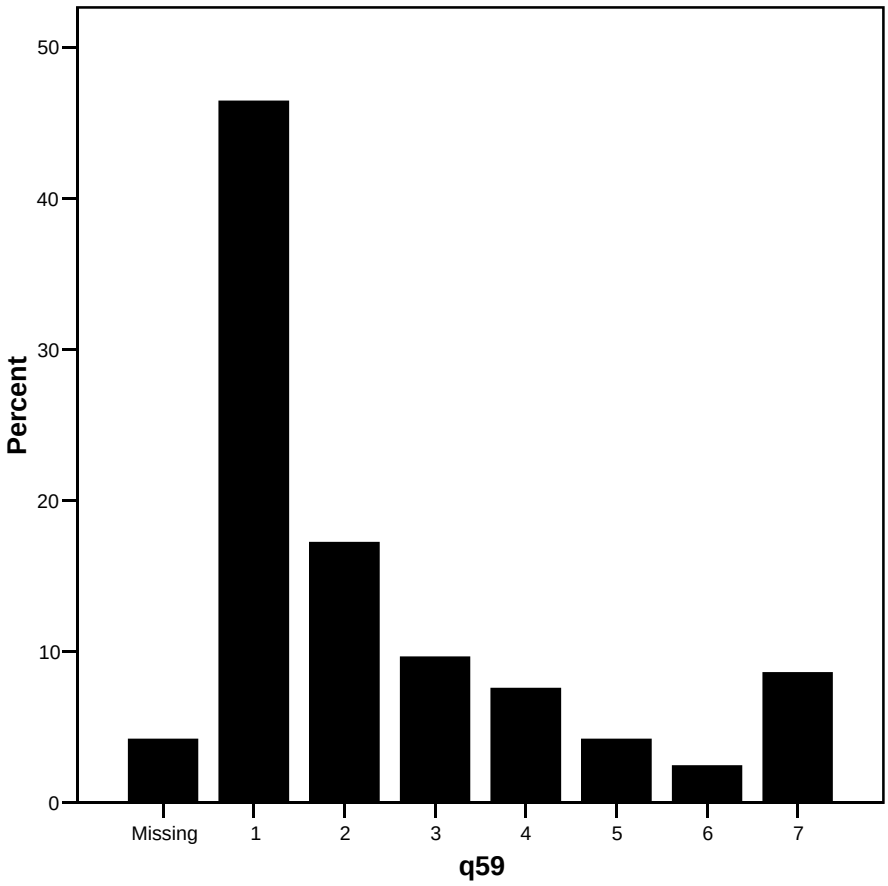

Figure 3. Distribution of response variable (Q59)

Q12 corresponds to the question "*During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club?*" It has 5 ordinal levels corresponding to: 0 days, 1 day, 2 or 3 days, 4 or 5 days, 6 or more days.

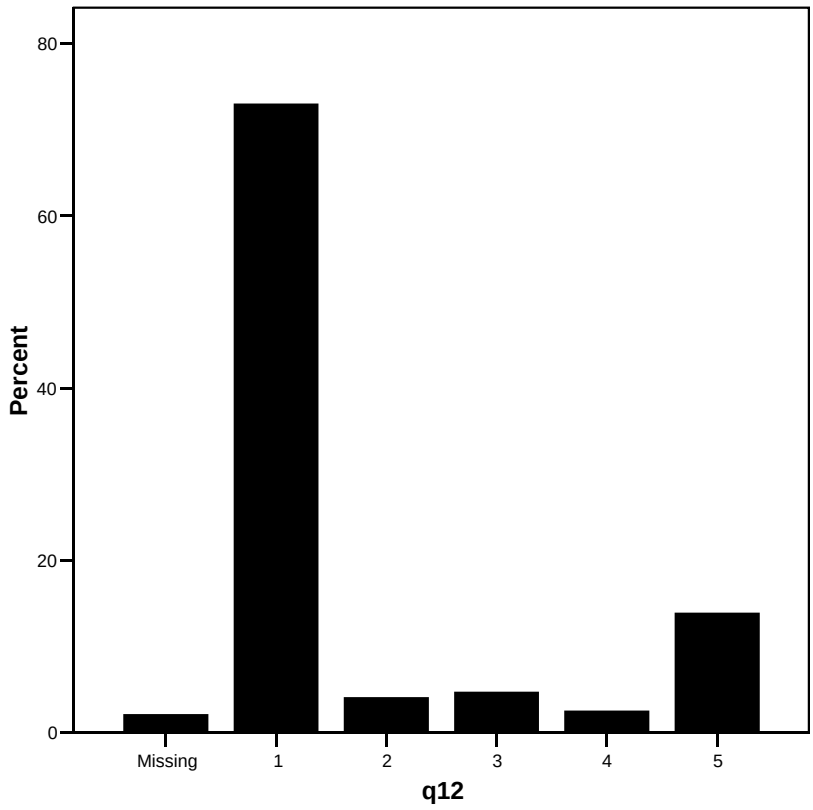

Figure 4. Distribution of response variable (Q12)

Q30 corresponds to the question "*During the past 30 days, on how many days did you smoke cigarettes?*" It has 7 ordinal levels corresponding to: 0 days, 1 or 2 days, 3 to 5 days, 6 to 9 days, 10 to 19 days, 20 to 29 days, and all 30 days.

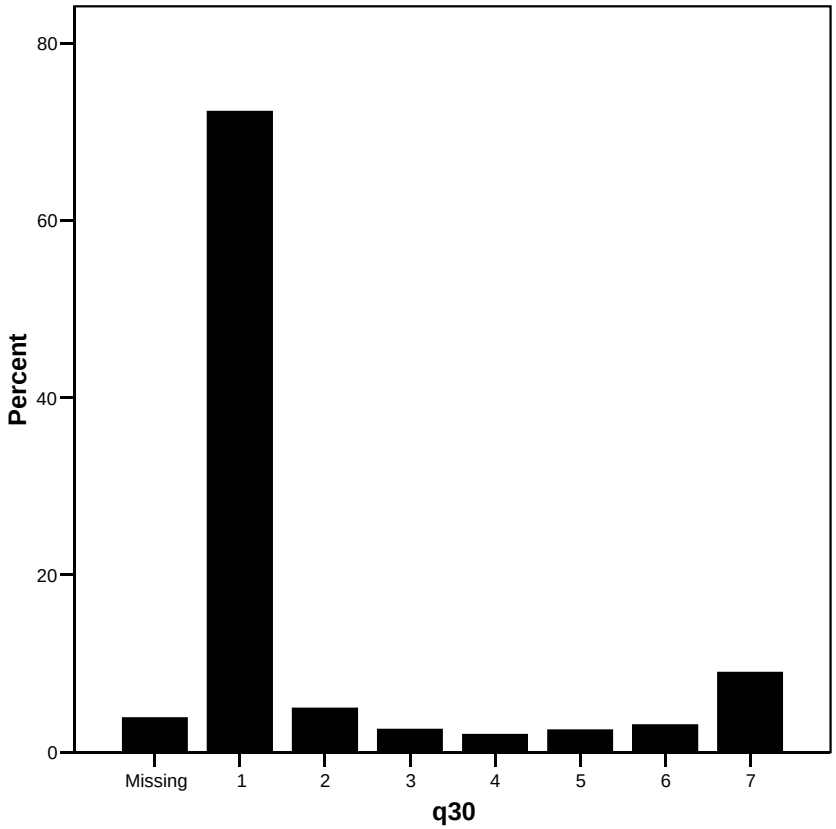

Figure 5. Distribution of response variable (Q30)

The percentage of missing data for Q44, Q59, Q12 and, Q30 were 1.2%, 4.2%, 2.1%, and 3.8% respectively. Because the percentage of missing data for ordinal predictors were less than 5%, it can be ignored without introducing serious bias (Roth, 1994). Our response variable had more than 10% data missing, hence, some form of missing data technique was required. Table 2 shows frequencies and percentages of study variables.

Table 2.

Frequencies and Percentages of study variables

	Ordinal Level	Frequency	Percentage (%)
Q39	1	373	24.2
	2	215	14.0
	3	244	15.8
	4	146	9.5
	5	159	10.3
	6	114	7.4
	7	134	8.7
	Missing	156	10.1
Q12	1	1124	72.9
	2	62	4.0
	3	72	4.7
	4	38	2.5
	5	213	13.8
	Missing	32	2.1
Q30	1	1114	72.3
	2	76	4.9
	3	39	2.5
	4	30	1.9
	5	38	2.5
	6	47	3.0
	7	138	9.0
	Missing	59	3.8
Q44	1	918	59.6
	2	165	10.7
	3	131	8.5
	4	64	4.2
	5	63	4.1
	6	49	3.2
	7	133	8.6
	Missing	18	1.2
Q59	1	715	46.4
	2	265	17.2
	3	148	9.6
	4	116	7.5
	5	64	4.2
	6	37	2.4
	7	132	8.6
	Missing	64	4.2

Assumption

MAR assumption was used in this paper. It means that "missingness" of a study variable is conditional on variables of similar nature. For the sake of imputation we assumed that the missing observations of our key variable (Q39) are conditional on our observed covariates (Q44, Q59, Q12, and Q30). These five variables were assumed to be sensitive or embarrassing for the high school children in Tennessee.

The main idea of this paper is to impute the missing values of Q39 using a cumulative logit model. In order to do this we first dichotomized our ordinal covariates (Q44, Q59, Q12, and Q30). For selecting the ideal cutpoints of our ordinal covariates, we dichotomized our response variable (Q39) with a never vs. ever routine. Because dichotomization of ordinal variables leads to loss of information, we dichotomized our ordinal predictors based on the principle of maximally selected chi-square statistics. The details of which are given in section 5.4. After finding the cutpoints for each of our ordinal covariates, we used these dichotomized ordinal predictors along with our seven leveled response variable (Q39) in a cumulative logit model routine to build our imputation framework. The details of our method are given in section 5.5. We will call this as method A. We also compared method A to two other MI based methods: IP (method B) and EMB (method C). The details of method B and C are given in section 5.6 and 5.7 respectively.

Maximally Selected Chi-square Statistics

The cut points for our ordinal predictors were based on maximally selected chi square statistics over all possible cutpoints. The likelihood-ratio chi-square statistic (G^2) [Wilks. S. S, 1935] involves the ratios between the observed and expected frequencies. The statistic is computed as follows:

$G^2 = 2\sum_i \sum_j n_{ij} \ln (n_{ij}/e_{ij})$

Where $\mathbf{n} = \sum_i \sum_j \mathbf{n_{ij}}$ is the overall total, $n_{i.}n_{.j}$ is the product of row total and column total and $e_{ij} = \frac{n_{i.}n_{.j}}{n}$

Let X be an ordinal variable with k distinct levels. X can be transformed into binary variables $X^{(k)}$ for $k = 1,\ldots,k\text{-}1$ as follows

$X^{(k)} = 1$ if $X \leq k$,

$X^{(k)} = 2$ if $X > k$,

For example $Q44^{(3)}$ represents the binary variable obtained by dichotomizing Q44 between ordinal level 3 and 4.

Given below are the G^2 values for all possible cutpoints. Tables 3, 4, 5, & 6 represents the cross tabulation of each ordinal predictor. The ideal cutpoint for each predictor is italicized.

Table 3.
Likelihood Ratio Chi-square Values for All Possible Cutpoints for Q44

$Q39_r$ × Predictor (Q44)		
Predictor	Likelihood χ^2 (G^2)	Degrees of freedom
Q44	300.5039	6
$Q44^{(1)}$	*283.5227*	1
$Q44^{(2)}$	230.7058	1
$Q44^{(3)}$	150.2163	1
$Q44^{(4)}$	118.3618	1
$Q44^{(5)}$	82.3443	1
$Q44^{(6)}$	52.4395	1

Table 4.
Likelihood Ratio Chi-square Values for All Possible Cutpoints for Q59

$Q39_r$ × Predictor (Q59)		
Predictor	Likelihood χ^2 (G^2)	Degrees of freedom
Q59	192.4136	6
$Q59^{(1)}$	*179.5074*	1
$Q59^{(2)}$	109.4441	1
$Q59^{(3)}$	88.0584	1
$Q59^{(4)}$	67.3443	1
$Q59^{(5)}$	58.6335	1
$Q59^{(6)}$	42.6095	1

Table 5.
Likelihood Ratio Chi-square Values for All Possible Cutpoints for Q12

$Q39_r$ × Predictor (Q12)		
Predictor	Likelihood χ^2 (G^2)	Degrees of freedom
Q12	53.0192	4
$Q12^{(1)}$	*44.3132*	1
$Q12^{(2)}$	29.4213	1
$Q12^{(3)}$	35.4224	1
$Q12^{(4)}$	29.6660	1

Table 6.
Likelihood Ratio Chi-square Values for All Possible Cutpoints for Q12

$Q39_r$ ×. Predictor (Q30)		
Predictor	Likelihood χ^2 (G^2)	Degrees of freedom
Q30	178.3255	6
$Q30^{(1)}$	*169.8321*	1
$Q30^{(2)}$	149.5152	1
$Q30^{(3)}$	125.9653	1
$Q30^{(4)}$	113.0258	1
$Q30^{(5)}$	88.7663	1
$Q30^{(6)}$	64.0518	1

In our case we found that the ideal cutpoints for all our ordinal variables were between ordinal level 1 and 2. But this may not always be the case, hence, the readers are requested to refer to "Intelligent dichotomies" presented by Drane at American Academy of Health Behavior, Savannah, GA, 2007. Attention must be paid while selecting the $X^{(k)}$ yielding the smallest p- value and declaring *k* as the ideal cutpoint of X because some people may argue that it may lead to inflation of type I error rate as it involves multiple

testing of many 2×2 tables (Altman et al, 1994). Betensky and Rabinowitz (1999) have investigated maximally selected χ^2 statistics in case of $K \times 2$ contingency tables. Koziol (1991) and Boulesteic x (2006) have derived the exact distribution of the maximally selected χ^2 statistics.

Method A

After dichotomizing our ordinal predictors with the routine given in section 5.4, we used the observed values of these dichotomized variables in building a multilevel cumulative logit model with Q39 as our response variable (the one that needs to be imputed). Individual predicted probabilities (IPs) at each response level (RL) were obtained for each observation. These IPs were averaged for each RL and the seven average IPs obtained were used in randomly imputing the missing values of Q39 using a uniform distribution. Computations in this section were done using SAS® (SAS Institute Inc., Cary, NC).

Method B

This method uses the Imputation-Posterior approach given by Schafer. It uses Markov Chain Monte Carlo (MCMC) method. The initial parameter estimates are obtained by running Expectation-Maximization algorithm (Rubin, 1976) until convergence is achieved (maximum iterations =1000). These EM estimates were used as starting values, 500 cycle of MCMC full data imputation were performed using a ridge prior. Only a single dataset was generated for the purpose of comparison. Please note that MCMC method requires a multivariate normal model but our variables were highly skewed. If the amount of missing information is not large, the MI based inferences are

robust to departure from multivariate normality (Schafer 1997, pp. 147 – 148). There are some specialized MCMC based imputation models for discrete variables but we have not included them in our study (Schafer, 1997). SAS® Proc MI was used for computation in this section.

Method C

A bootstrapping based algorithm (EMB) is used in this method. We chose this method because it's fast and easy to use. Instead of using draws from posterior distribution, this method uses sampling with replacement. A fast EM algorithm is run on each sample. For each set of estimates, the original sample units are used to impute the incomplete observation. Again, only a single dataset was imputed for the purpose of comparison. There has been some evidence that EBM works well with discrete variables (King et al., 2001). Ameliaview, a standalone, GUI software was used to run the EMB algorithm in this section. Figure 6 depicts the density of imputed and observed data using EMB algorithm.

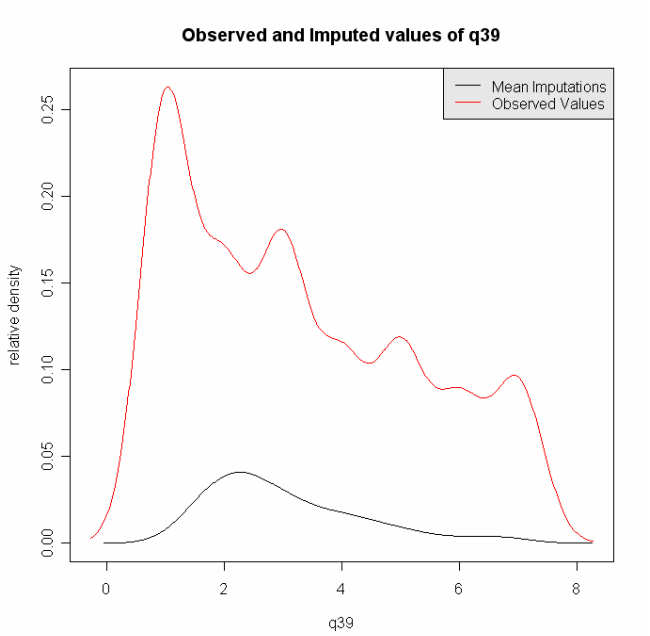

Figure 6.
Density of imputed and observed data. Imputed values are captured within the bounds of observed data but it poorly follows the observed distribution

CHAPTER 6

RESULTS

The response variable (Q39) was imputed using three methods (method A, B, and C). Table 7 shows the frequency and percentages of Q39 after imputation by method A, B, and C.

Table 7.

Frequencies and percentages of response variable Q39 after imputation by method A, B, and C

	Method A		Method B		Method C	
Q39	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
1	409	26.6	397	25.8	415	26.9
2	241	15.6	238	15.5	239	15.5
3	276	17.9	288	18.7	270	17.5
4	162	10.5	180	11.7	164	10.6
5	184	11.9	177	11.5	187	12.1
6	127	8.2	122	7.9	122	7.9
7	141	9.1	138	9.0	143	9.3
Total	1540	100	1540	100	1540	100

The distribution of complete cases (CC) for Q39 was compared with the imputed distributions from the three methods, using Minimum Discriminant Information statistic (MDIS). It is also the likelihood ratio chi square statistics (G^2) and is given by the formula:

$$G^2 = 2n\sum_{i=1}^{7} p_i \log\frac{p_i}{p_{iI}}$$ with degrees of freedom (df) = 6

Here n = Total number of cases, p_i is the observed probability of response variable (Q39) at i^{th} response level and p_{iI} is the imputed probability of response variable at i^{th} response level where $i = 1,\ldots, 7$

G^2 obtained from each method is given in table B. Although all three methods performed well, our method yielded the least G^2 statistics, which indicates that the distribution obtained after imputation is highly similar to the distribution of observed cases (CC).

Table 8.
Likelihood ratio chi square statistics for method A, B, and C

Method	Likelihood ratio Chi Square Statistics (df)
Method A	1.09981 (6)
Method B	5.86435 (6)
Method C	2.92884 (6)

df: Degrees of freedom

CHAPTER 7

DISCUSSION

Missing data are mostly unavoidable and generally occurs for unknown reasons. The unknown mechanisms behind "missingness" can only be assumed. These assumptions are usually untestable. For meaningful statistical estimation, this assumption should closely correspond to the real mechanism behind missing data. Ultimate users of dataset usually don't have the knowledge and expertise to handle missing data, hence, the database constructors, who typically know more about the reasons for missingness, should be responsible for modeling the missing data (Rubin, 1996).

Our method of imputation involves building a cumulative logit model. This is the most crucial step in our suggested method. The intention here is neither to build a parsimonious model, nor to describe a causal relationship among variables and, hence, recognition of dependent and independent variables is not important at this stage, although an attempt should be made to preserve the effects of interests. The chances of meeting the MAR assumption increases as we increase the number of variables. By using Q44, Q59, Q12, and Q30 as covariates we don't necessarily mean that these variables have a causal relationship with Q39. We want to emphasize that the idea behind building this pre-imputation model is to meet the MAR assumption without significantly distorting the effects of interest in post-imputation analysis.

This paper also emphasizes a statistically sound but infrequently used method of ideal cut-point selection for ordinal variables. Instead of randomly selecting a cut-point of an ordinal variable, we emphasize using the maximally selected chi square statistics. It is

well known that dichotomization of variables lead to loss of information. A cut-point based on maximal Likelihood-ratio chi-square statistic (G^2) will lead to minimal loss of information.

Finally our paper compares the performance of our method of imputation with two other well known methods of imputation, namely: 1) Markov-chain, Monte Carlo based Imputation-Posterior (IP) algorithm [method B] and 2) Bootstrapping based EMB algorithm (method C). Minimum Discriminant Information statistic (MDIS) based comparison of these three methods showed that our method performed reasonably well in comparison to the other two methods.

However, there are several limitations in our paper. Firstly, our cumulative logit model allows imputation of one variable at a time, which can be time-consuming. Secondly, if there is significant amount of missingness in the independent variables, our model will either be inappropriate or may need some adjustments which again can be complicated. Thirdly, performance of our model at larger amounts of missingness (e.g. 20%, 30%, 50%, etc) has not been tested, hence, further research needs to be done before this model can be applicable for general use. One prospective way to do this is to simulate various levels of missingness and compare imputed data with complete data.

A subjective tabulation of some of the common missing data techniques (MDT) with respect to their ease of use and validity is given in Table 9.
These have been ranked on a scale of 1 to 10 (For "Ease of use": 1 represents easiest to use, and 10 represents most difficult to use and for "Validity of parameter estimates": 1 represents most valid estimates, 10 represents least valid estimates).

Table 9.

Subjective tabulation of some of the common missing data techniques with respect to their ease of use and validity

Missing Data Technique.	**Ease of use.**	**Validity of parameter estimates.**
Deletion	1	9
Mean Substitution	2	8
Random Imputation from observed data	4	7
Covariate based prediction.	4	4
Drawing missing values from its predictive distribution under a specified model	8	3
Multiple Imputation	9	2

In the real world, covariate based prediction of missing values seems to be a good choice for end-users using publicly shared datasets with varying degree of computing knowledge and statistical expertise.

REFERENCES

Allison, P.D. (2002) *Missing data.* Thousand Oaks, CA: Sage

Altman, D.G., Lausen, B., & Sauerbrei, W. (1994) Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute, 86, 11*, 829–835.

Betensky, R.A., & Rabinowitz, D. (1999). Maximally selected χ^2 statistics for k × 2 tables. *Biometrics 55,* 317–320.

Boulesteix, A. (2006). Maximally Selected Chi-square Statistics for Ordinal Variables. *Biometrical Journal 48*, 3, 451–462.

Chen, L., Drane, M.T., Valois. R.F., & Drane. J.W. (2005). Multiple imputation for missing ordinal data. *Journal of Modern Applied Statistical Methods. Vol. 4, No.1*, 288-299.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Ford, B.L. (1983). An overview of Hot-deck procedures. In Madow W.G, Olkin. I., & Rubin. D. B. (Eds.), Incomplete data in sample surveys. Volume II: Theories and bibliographies, 185-207. New York: Academic Press

Graham, J.W., Hofer, S.M., & Piccinin, A.M. (1994). *Analysis with missing data in drug prevention research*. In Collins, L. & Seitz, L. (Eds.), National Institute on Drug Abuse Monograph Series, Vol. 142, pp.13-62. Washington, D.C: National Institute on Drug Abuse.

Honaker, J., King, G., & Blackwell, M. (2006), Amelia Software. Retrieved December 15, 2006, from the Web site: *http://gking.harvard.edu/amelia*.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., & Herring, A.H. (2005). Missing-Data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100, 332–346.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001), Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, *95*, 49–69.

Kolbe, L.J. (1990). An epidemiologic surveillance system to monitor the prevalence of youth behaviors that most affect health. *Journal of Health Education*, *21, 6,* 44-48.

Koziol, J.A. (1991). On maximally selected Chi-square statistics. *Biometrics 47*, 1557–

1561.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: J. Wiley & Sons.

Little, R.J.A., & Rubin, D.B. (1989). The Analysis of social science data with missing values. *Sociological Methods & Research, 18*, 292-326.

McLachlan, G.J., & Krishnan, T. (1996). *The EM algorithm and extensions.* New York: Wiley.

Rizvi, M.H. (1983). Hot deck procedures: Introduction. In Madow W. G, Olkin I (Eds.), Incomplete data in sample surveys. Volume III: Proceedings of the symposium, (pp.351-352). New York: Academic Press.

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology,* 47, 537-560.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: J. Wiley & Sons.

Rubin, D.B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association, 91*, 473-489.

SAS Institute Inc. 2004. *What's New in SAS® 9.0, 9.1, 9.1.2, and 9.1.3*. Cary, NC: SAS Institute.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data.* London: Chapman & Hall/CRC

Shah, B.V., Barnwell, G.B., & Bieler, G.S. (1997). *SUDAAN*, software for the statistical analysis of correlated data, User's Manual. Release 7.5 ed. Research Triangle Park, NC: Research Triangle Institute.

Tennessee Department of Education website Retrieved March 25, 2007, from the Web site: *http://www.k-12.state.tn.us/yrbs/ciyrbs05/index.htm*

Van Buuren, S., Boshuizen, H.C., & Knook, D.L.(1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

Van Buuren, S., Brand, J.P.L., Groothius-Oudshoorn, C.G.M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064.

Wilks, S.S. (1935). The likelihood test of independence in Contingency tables. *The Annals of Mathematical Statistics*, *6,* 190-196.

VITA

ANDALEEB A. AHMED, MD., MPH.

Personal Data: Date of Birth: June 06, 1978
Nationality: Indian
Gender: Male
Marital Status: Single

Education: University College of Medical Sciences, University of Delhi, Delhi, India.
Bachelor of Medicine and Bachelor of Surgery, 2004

East Tennessee State University, College of Public and Allied Health, Johnson City, Tennessee.
Master of Public Health, Biostatistics, 2006-2007

Good Samaritan Hospital of Maryland / Johns Hopkins Hospital, Baltimore, Maryland.
Internal Medicine, Graduate Medical Education, 2007-2010

Work Experience: G.T.B. Hospital, Delhi, India.
Clinical Internship, 2003

University of Alabama at Birmingham, Department of Anesthesiology, Birmingham, Alabama.
Research Assistant, 2005

University of Alabama at Birmingham, Center for Health Promotion, Birmingham, Alabama.
Research Assistant, 2006

University of Alabama at Birmingham, General Clinic Research Center, Birmingham, Alabama.
Research Assistant, 2006

East Tennessee State University, Department of Public Health, Johnson City, Tennessee.
Graduate Assistant, 2006-2007