

**University of Illinois at Chicago
School of Public Health
Division of Epidemiology and Biostatistics**

*Technical report#:2007-004
June 2007*

Title: A distance-based rounding strategy for post-imputation
ordinal data

Author: Hakan Demirtas

**Affiliation: University of Illinois at Chicago, Division of Epidemiology and
Biostatistics.**

A distance-based rounding strategy for post-imputation ordinal data

Hakan Demirtas*

June 19, 2007

Abstract

Multiple imputation has emerged as a widely used model-based approach in dealing with incomplete data in many applications areas. Gaussian and log-linear imputation models are fairly straightforward to implement for continuous and discrete data, respectively. However, in missing data settings that include a mix of continuous and discrete variables, correct specification of the imputation model could be a daunting task due to the lack of flexible models for the joint distribution of variables of different nature. This complication, along with accessibility to software packages that are capable of carrying out multiple imputation under the assumption of joint multivariate normality, appears to encourage applied researchers for pragmatically treating the discrete variables as continuous for imputation purposes, and subsequently rounding the imputed values to the nearest observed category. In this article, we introduce a distance-based rounding approach for ordinal variables in the presence of continuous ones. The first step of the proposed rounding process is predicated upon creating indicator variables that correspond to the ordinal levels, followed by jointly imputing all variables under the assumption of multivariate normality. The set of imputed values are then converted to the ordinal scale based on their Euclidean distances to the set of indicators, with minimal distance corresponding to the closest match. We compare the performance of this technique to simple rounding in terms of commonly accepted accuracy and precision measures with simulated data sets that are generated around a real data set from psychiatric research.

Key Words: Multiple imputation; Rounding; Bias; Precision

1 Introduction

Missing data is a commonly occurring phenomenon in biopharmaceutical practice. Missing values generally complicate the statistical analysis in terms of reduced power, degraded

*Hakan Demirtas (e-mail:demirtas@uic.edu) is an Assistant Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL, 60612.

confidence intervals, and biased parameter estimates, leading to false inferences (Little and Rubin, 2002).

Multiple imputation (MI) has been an increasingly popular model-based simulation technique in the analysis of missing data. MI replaces each missing datum with a set of plausible values that are drawn from a predictive distribution. Key ideas and advantages of MI are reviewed by Rubin (1996, 2004) and Schafer (1997, 1999). For an extensive bibliography and a software review, see Rubin (1996) and Horton and Lipsitz (2001), respectively. MI allows researchers to use more conventional models and software; an imputed data set may be analyzed by literally any method that would be suitable if the data were complete. As computing environments and statistical models grow increasingly sophisticated, the value of using familiar methods and software becomes important. Furthermore, there are still many classes of problems for which no direct maximum likelihood procedure is available. Even when such a procedure exists, MI can be more attractive due to fact that the separation of the imputation phase from the analysis phase lends greater flexibility to the entire process. Lastly, MI singles out missing data as a source of random variation distinct from ordinary sampling variability.

The fundamental step in parametric MI is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data which usually entails positing a model for the data and using it to derive this conditional distribution. For continuous data, joint multivariate normality among the variables has often been perceived as a natural assumption since the conditional distribution of the missing data given the observed data is then also multivariate normal. When all the variables are categorical, a log-linear imputation model can be used (Schafer, 1997). If the sample size is assumed fixed, the set of cell frequencies in a contingency table has a multinomial distribution. If there are no restrictions on the parameters other than they are true probabilities, then the model is said to be saturated. Log-linear models are a flexible class of models for specifying possible dependencies among variables. With complete data, using a Dirichlet prior distribution for the saturated model leads to a conjugate analysis. The posterior distribution is again Dirichlet with updated parameters involving the data and prior parameters.

The motivation of this article is two-fold. First, when the data only consist of categorical variables, a saturated model with a Dirichlet prior is the most commonly employed imputation method in practice. This approach has been shown to work in some settings; it is quite general in the sense that it allows for three-way and higher-order associations

among variables. However, some of these associations may be poorly estimated in many applications, because the observed data may be sparse and may not be able to support that complexity. Even when higher-order terms can be eliminated, imputation with the log-linear modeling framework can be computationally very demanding due to massive proliferation of model parameters as the number of missing data patterns grow. Second, most real data sets consists of a mix of continuous and discrete variables. Although the general location model with conditional Gaussian structure that is designed to work for variables of mixed nature has its merits, the joint distribution that leads to the presumably correct imputation model is not always straightforward to formulate. The limitations of conditional Gaussian model have been illustrated by Belin et al. (1999).

In such situations, practitioners generally resort to continuous imputation techniques by treating discrete variables as continuous, and subsequently rounding the imputed values to the nearest observed category. The central focus of this work is comparing simple rounding in which incomplete ordinal variables that are regarded as continuous for the purpose of imputation, then rounded to the nearest observed category after performing MI, and a distance-based rounding approach whose essence is creating indicator variables that correspond to the ordinal levels, followed by jointly imputing all variables under the assumption of multivariate normality. The set of imputed values are then converted to the ordinal scale based on their Euclidean distances to the set of indicators, with minimal distance corresponding to the closest match. To the best of our knowledge, rounding for post-imputation ordinal data has not yet been investigated; and the motivation of this manuscript is to fill this gap.

The organization of this article is as follows: In the next section, we elaborate on the rounding method we propose. In Section 3, we give a brief overview on the essentials of MI under normal models. In Section 4, we describe a simulation study where incomplete simulated data sets were generated around a real data set from psychiatric research. We implement the two rounding strategies under consideration on imputed data sets and evaluate the comparative performances in terms of bias and efficiency properties for the population parameters we have chosen. Section 5 includes discussion and concluding remarks.

2 Proposed rounding approach

We present the proposed approach for a single incomplete ordinal variable that has k levels. It is well-known that a set of $k - 1$ linearly independent indicator variables contains all the relevant information. Let Z_{ind} be the associated indicator matrix of dimension $k \times (k - 1)$, whose elements are Z_{ij} , where $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k - 1$. $Z_{ij} = 1$ when $i + j = k + 1$, and 0 otherwise. The vector for each i , $Z_i = (Z_{i1}, Z_{i2}, \dots, Z_{i,k-1})^T$. Clearly, if an observation is missing, all newly created $k - 1$ variables are coded as “missing”. After MI is performed on all available complete or incomplete variables in the system including Z_i 's under the assumption of joint multivariate normality whose operational details are given in Section 3, we obtain $k - 1$ continuous values for each of the missing ordinal level. Let $W = (W_1, \dots, W_{k-1})^T$ be the vector that contains imputed values. Then, via the Euclidean distance between W and Z_i , $\sqrt{\sum_{j=1}^{k-1} (W_j - Z_{ij})^2}$, that are calculated for $i = 1, 2, \dots, k$, the closest ordinal category can be determined by the minimum of these k distances. The same process applies to all rows of the data that have unobserved observations for the ordinal variable under consideration.

For illustration purposes, let us assume that $k = 4$, and the ordinal levels are 1, 2, 3, 4. Then,

$$Z_{ind} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

with rows corresponding to 1, 2, 3, 4 in order. Suppose the set of imputed values are $(0.2, 0.7, 0.6)$. The Euclidean distances with respect to 1, 2, 3, 4 are, 0.94, 0.83, 0.7, 1.22, respectively. The minimum distance matches with the category 3. Extending this methodology to multiple ordinal variables is fairly straightforward in the sense that conversion to the original ordinal scale can be done on a variable by variable basis.

Both the simple and proposed rounding methods rely on imputing under the normality assumption. In the next section, we describe the most salient characteristics of normal imputation models.

3 Imputing under normal models

Let y_{ij} denote an individual element of $Y = (Y_{obs}, Y_{mis})$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$, where Y , Y_{obs} , and Y_{mis} denote complete data, observed data, and missing data, respectively. The i^{th} row of Y is $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^T$. Assume that y_1, y_2, \dots, y_n are independent realizations of a random vector, denoted as (Y_1, Y_2, \dots, Y_p) , which has a multivariate normal distribution with mean vector μ and covariance matrix Σ ; that is $y_1, y_2, \dots, y_n | \theta \sim N(\mu, \Sigma)$, where $\theta = (\mu, \Sigma)$ is the unknown parameter and Σ is positive definite. When imputations are created under Bayesian arguments, MI has a natural interpretation as an approximate Bayesian inference for the quantities of interest based on the observed data. MI can be performed by first running the EM algorithm, and then by employing a data augmentation procedure, as implemented in some software packages. The EM algorithm is useful for two reasons: it provides good starting values for the data augmentation scheme, and it gives us an idea about the convergence behavior. Data augmentation using the Bayesian paradigm has been perceived as a natural tool to create multiply imputed data sets. For further details, see Schafer (1997) and Schimert et al. (2001). When both μ and Σ are unknown, the conjugate class for the multivariate normal data model is the normal inverted-Wishart family. When no strong prior information is available about θ , one may apply Bayes' theorem with the improper prior. In the simulated examples, a noninformative prior was used to reflect a state of relative ignorance.

Initial estimates for θ are usually obtained by the EM algorithm. Then, data augmentation scheme is implemented as follows: First, a value of missing data from the conditional predictive distribution of Y_{mis} , $Y_{mis}^{(t+1)} \sim P(Y_{mis} | Y_{obs}, \theta^{(t)})$, is drawn. Then, conditioning on $Y_{mis}^{(t+1)}$, a new value of θ from its complete-data posterior, $\theta^{(t+1)} \sim P(\theta | Y_{obs}, Y_{mis}^{(t+1)})$ is drawn. Repeating these two steps from a starting value $\theta^{(0)}$ yields a stochastic sequence $(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, \dots$ whose stationary distribution is $P(\theta, Y_{mis} | Y_{obs})$, and the subsequences $\theta^{(t)}$ and $Y_{mis}^{(t)}$ have $P(\theta | Y_{obs})$ and $P(Y_{mis} | Y_{obs})$ as their respective distributions. For a reasonably large number of iterations, the convergence to these stationary distributions is achieved. Since the complete-data likelihood is assumed to follow a multivariate normal distribution, drawing from conditional distributions above is relatively straightforward and can be performed by applying sweep operators to subsets of the vector μ and the matrix Σ .

4 A simulation study devised around real data

Describing a real phenomenon by generating an environment within which the process of interest is assumed to operate is not uncommon and is often the only feasible way of evaluation. The idea of creating many imperfect proxies of what is believed to be the truth is based on simulating the performance of a method by proposing a variety of populations and missingness mechanisms capable of producing data like those actually seen; then assessing the behavior of various methods over repeated samples from each population, and identifying methods that seem to perform well for a variety of populations. In this section, we present a simulation study driven by this approach to evaluate the differential performances under the two competing rounding rules for ordinal data.

Our real-data example that anchors the simulation study comes from Hedeker and Gibbons (1997) who use the data from the National Institute of Mental Health Schizophrenia Collaborative Study. Patients were randomly assigned to receive one of three anti-psychotic medications or a placebo. We collapsed the subjects from the three drug treatments into a single group, because the performance of the three drugs was reported to be quite similar (Hedeker and Gibbons, 1997). The outcome of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which we treat as continuous. Of note, there are non-integer values due to multiple raters in the data set. Measurements were planned for weeks 0, 1, 3, and 6, but missing values occurred primarily due to drop-out. A few subjects had missing measurements and subsequently returned; for simplicity we have removed these. (We could have included these cases with non-monotone missingness, as Hedeker and Gibbons (1997) did. We decided to exclude them to simplify the task of constructing alternative hypothetical population models for our simulations.) The monotone missingness assumption (drop-out) has no bearing on the conclusions drawn in this paper and was merely done for convenience. A small number of measurements were also taken at intermediate time points (weeks 2, 4, and 5) which we also ignore. With these exclusions, the sample contains 312 patients who received a drug and 101 who received a placebo. In the drug group, 3 patients dropped out immediately after week 0, 27 dropped out after week 1, 34 dropped out after week 3, and 248 completed the study. In the placebo group, no patients dropped out after week 0, 18 dropped out after week 1, 19 dropped out after week 3, and there were 64 completers. Overall completion rate is about 75%. Hedeker and Gibbons (1997) noted that the mean response profiles are approximately linear when

plotted against the square root of week, and they express time on the square-root scale in their models. Adopting this convention, we define time to be the square root of week. This data were previously analyzed by Demirtas and Schafer (2003), Demirtas (2005a, 2005b) in the context of pattern-mixture modeling.

4.1 Complete data generation

We generate the complete data based on well-known linear mixed-effects model (Laird and Ware, 1982). Let $y_i = (y_{i1}, \dots, y_{in_i})^T$ denote the responses for subject i . The model is

$$y_i = X_i\beta + Z_ib_i + \epsilon_i, \tag{1}$$

where X_i ($n_i \times p$) and Z_i ($n_i \times q$) contain covariates, β contains fixed effects, $b_i \sim N(0, \psi)$ contains random effects, and $\epsilon_i \sim N(0, \sigma^2 V_i)$. Times of measurement are often incorporated into X_i and Z_i , allowing the response trajectories to vary by subject. Common choices for V_i include the identity or patterned (e.g. autoregressive or banded) matrices that reflect serial correlation. In this specific example, y_i 's are the responses for individual i at weeks 0, 1, 3, and 6. In our simulated population, we assume that $y_i = X_i\beta + Z_ib_i + \epsilon_i$ where the columns of X_i are a constant (one); G (0 for placebo, 1 for drug); T (square root of week); and GT . The columns of Z_i are a constant and T . The fixed effects are set to $\beta = (5.36, 0.05, -0.32, -0.65)^T$, the random effects b_i are normally distributed with covariance matrix

$$\psi = \begin{bmatrix} 0.35 & 0.04 \\ 0.04 & 0.23 \end{bmatrix},$$

and the elements of ϵ_i are independent and normal with variance $\sigma^2 = 0.60$. The number of subjects n was chosen to be 413 as in the original data set.

4.2 Missingness mechanism

We assume that drop-out occurs by the following selection process: the probability that patient i drops out immediately *after* week $w = 0, 1, 3$ is

$$\text{expit}(\alpha_w + \gamma_1 y_{iw} + \gamma_2 y_{iw}^2 + \gamma_3 G),$$

where $\alpha_0 = -0.69$, $\alpha_1 = 2.27$, $\alpha_3 = 2.48$, $\gamma_1 = -2.02$, $\gamma_2 = 0.24$, and $\gamma_3 = -0.87$. Here, $\text{expit}(x) = 1/(1 + e^{-x})$. With this nonresponse mechanism, missing values are assumed to

be missing at random (MAR) in the sense defined by Rubin (1976); in the current work we are not concerned with departures from MAR.

The complete data generation process and missing-data mechanism presented herein, yield simulated trajectories that closely resemble the real data trends on average.

4.3 Imputation and rounding

We assume two imputation models for comparison purposes for each of the incomplete data sets generated. The first one is the normal model, where we create imputations following the standard approach of using a Bayesian predictive model of the missing data given the observed data (Schafer, 1997). The ordinal variable (severity of illness) was treated as continuous and taken as one of the components of the multivariate normal distribution for imputation. Next, imputed ordinal variable was rounded to the nearest observed category. In the suggested distance-based approach, three indicator variables were created. When the original ordinal variable is missing, all three indicators were coded as “NA”. We proceeded with MI under multivariate normality, and identified the “closest” ordinal category via Euclidean distances as articulated in Section 2. We ordinalized the response at the last time point with the thresholds that appear on the seminal paper of Hedeker and Gibbons (1994). The reason we do this only for the last observation is to illustrate our method in the presence of continuous variables in the spirit of better reality compatibility.

Let Y_4 be the response at the time of last measurement (week 6). In the version 1 (Hedeker and Gibbons, 1994), $Y_4 = I[1, 2.5) * 1 + I[2.5, 4.5) * 2 + I[4.5, 5.5) * 3 + I[5.5, 7] * 4$, where I denotes the indicator function. $[$ or $]$, and $($ or $)$ stand for inclusive and exclusive bounds of the indicator function, respectively. The true proportions with these thresholds are $(0.3019976, 0.4357385, 0.1478475, 0.1144165)$. In the version 2, $Y_4 = I[1, 2.5) * 1 + I[2.5, 3) * 2 + I[3, 4) * 3 + I[4, 7] * 4$, leading to the true proportion values of $(0.3019976, 0.1091671, 0.2270266, 0.3618717)$. In the version 1, an intermediate category, 2, is the dominant one, whereas in the version 2 the first and last categories are represented by the highest frequencies.

4.4 Parameters of interest

We compared the relative performances of simple and distance-based rounding on the proportion of the four ordinal categories.

4.5 Evaluation criteria

The simulation experiment was repeated $N = 1000$ times for each of the two scenarios. Evaluation is conducted via four quantities: a) *Standardized bias (SB)* is the relative magnitude of the raw bias to the overall uncertainty in the system. If the parameter of interest is θ , the standardized bias is $100 \times \frac{E(\hat{\theta}) - \theta}{SE(\hat{\theta})}$, where SE stands for standard error. If the standardized bias exceeds 50% in a positive or negative direction, then the bias begins to have a noticeable adverse impact on efficiency, coverage and error rates (Demirtas, 2004). b) *Coverage rate (CR)* is the percentage of times that the true parameter value is covered in the confidence interval. If a procedure is working well, the actual coverage should be close to the nominal rate (i.e. Type I error rates are properly controlled). We regard the performance of the interval procedure to be poor if its coverage drops below 90% (Collins et al., 2001). c) *Root-mean-square error (RMSE)* is a combined measure of bias and variance. It is considered to be arguably the best criterion for evaluating $\hat{\theta}$ in terms of combined accuracy and precision. $RMSE(\hat{\theta})$ is defined as $\sqrt{E_{\theta}[(\hat{\theta} - \theta)^2]}$. d) *Average width of confidence interval (AW)* is the distance between average lower and upper limits across 1000 confidence intervals. A high coverage rate along with narrow, calibrated confidence intervals translates into greater accuracy and higher power.

Under this specification, SB is a pure accuracy measure, AW is a pure efficiency measure, and CR and $RMSE$ are the hybrid measures of accuracy and precision. For more detailed discussion on this evaluation system, Demirtas (2007), Demirtas and Hedeker (2007), Demirtas, Freels and Yucel (2007), and Demirtas et al. (2007). Software implementation was done in R language (Version 2.4.1, 2007).

4.6 Results

In what follows, *SIMPLE* and *DISTANCE* stand for simple rounding and distance-based rounding. The resulting performances of the two rounding processes under the normal imputation model for the proportion of the four observed ordinal categories (p_1, p_2, p_3, p_4) in the first version of ordinal collapsing, are tabulated in Table 1. TV , AE , SB , $RMSE$, CR , and AW represent true value, average estimate, standardized bias, root-mean-square error, coverage rate, and average width, respectively. The number of simulation replicates, N is 1000; the length of the complete data vector, n is 413 as in the original data set. The number of significant digits varies depending on the quantity. SB and CR values that do

not fall in the acceptable range are shown with bold characters. The message in Table 1 is somewhat mixed. *SIMPLE* yields unacceptable biases in three out of four cases, and *DISTANCE* fails once, with one poor performance for the coverage rate in both. Absolute biases, *RMSE*'s, and average widths seem to have a 2 – 2 split. Although there is no clear pattern, *DISTANCE* appear to have a slight edge over *SIMPLE*. This is a bit surprising because the mode occurs in one of the middle categories, we would expect *SIMPLE* to deliver a better performance.

Table 2 follows the same structural format and shows the results for the second version of the ordinalization. As apparent from Table 2, *DISTANCE* is the obvious winner in this situation, both marginally and comparatively in both bias and coverage metrics. *SIMPLE* produces estimates that are biased for all parameters and coverage rates are unacceptably low in three cases, whereas *DISTANCE*'s performance appears satisfactory for all parameters. It is interesting to note that average widths are comparable, suggesting that rather drastic accuracy differences between *SIMPLE* and *DISTANCE* seems to disappear. This implies that the precision under the two rounding rules are comparable, which is fairly remarkable considering the historical trade-off between accuracy and efficiency.

Table 1 and 2 go here

5 Discussion

There are a few issues that deserve discussion. First, nonignorable modeling is beyond the scope of this manuscript. An assumption of MAR is commonly employed for MI. However, the theory of MI does not necessarily require MAR, MI may also be performed under non-ignorable models. For the purposes of this article, ignorable nonresponse is assumed; we were not concerned with departures from ignorability. Our focus was limited to rounding ordinal variables in a sensible way. Second, one may argue that generalizability of the simulation results is doubtful given countless number of scenarios that can be encountered in practice. Another possible objection may be that the simulated examples are insufficiently complex relative to most real-life applications. There is no way to contradict these statements; however, our intention was giving pragmatic advice to applied researchers who are accustomed to imputing with flexible and easy-to-implement software that are designed for continuous data, yet who have to deal with incomplete ordinal variables in their data. Furthermore, we do not take a position of advocacy for the proposed rounding method,

and are not presenting it as a miraculous approach that works in every imaginable situation. Nevertheless, it is worth to have this technique in the practitioners' toolbox; and this real-data driven simulated examples should unravel some aspects of rounding in MI. In addition, given sufficient time and resources, one can develop potentially better rules. Examples include other distance measures such as sum of absolute deviations; using the distances in this work, but employing a multinomial sampling where probabilities of selection are obtained through weighted inverses of the Euclidean distances; an ordinal logistic regression type of model where we first do simple rounding and treat this variable as the response and others as predictors, then align the rounded variable with respect to how well other variables predict them. We hope that this study serves as a milestone for further refinements.

REFERENCES

Belin, T.R., Hu M.Y., Young, A.S., Grusky O. (1999). Performance of a general location model with an ignorable missing data assumption in a multivariate mental health services study. *Statistics in Medicine*, 18:3123–3135.

Collins, L.M., Schafer, J.L., Kam, C.H. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6:330–351.

Demirtas, H. (2004). Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*, 58:466–482.

Demirtas, H. (2005a). Bayesian analysis of hierarchical pattern-mixture models for clinical trials data with attrition and comparisons to commonly used ad-hoc and model-based approaches. *Journal of Biopharmaceutical Statistics*, 25:383–402.

Demirtas, H. (2005b). Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 24: 2345–2363.

Demirtas, H. (2007). Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds. *Communications in Statistics-Simulation and Computation* (in press).

Demirtas, H., Arguelles, L.M., Chung, H., Hedeker, D. (2007). On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Computational Statistics and Data Analysis*, 51:4064–4068.

Demirtas, H., Freels, S.A., Yucel, R.M. (2007). Plausibility of multivariate normal-

ity assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation* (in press).

Demirtas, H., Hedeker, D. (2007). Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine*, 26:782–799.

Demirtas, H., Schafer, J.L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22:2253–2575.

Hedeker, D., Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50:933–944.

Hedeker, D., Gibbons, R.D. (1997). Application of random effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2:64–78.

Horton, J.H., Lipsitz, S.R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55:244–254.

Laird, N.M., Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.

R Development Core Team. (2007). *R: A language and environment for statistical computing, Version 2.4.1*. Vienna, Austria. URL: <http://www.r-project.org>.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 21:581–592.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91:473–520.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley Classic Library.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15.

Schimert, J., Schafer, J.L., Hesterberg, T., Fraley, C., Clarkson, D.B. (2001). *Analyzing Data with Missing Values in S-plus*. Seattle, WA: Data Analysis Products Division, Insightful Corp.

Table 1: The performance of the two rounding processes under the normal imputation model for the proportion of the four observed ordinal categories in the first version of ordinal collapsing. *SIMPLE*, *DISTANCE*, *TV*, *AE*, *SB*, *RMSE*, *CR*, and *AW* stand for simple rounding, distance-based rounding, true value, average estimate, standardized bias, root-mean-square error, coverage rate, and average width, respectively. The number of simulation replicates, N is 1000; the length of the complete data vector, n is 413 as in the original data set.

Rounding	Parameter	<i>TV</i>	<i>AE</i>	<i>SB</i>	<i>RMSE</i>	<i>CR</i>	<i>AW</i>
<i>SIMPLE</i>	p_1	0.3019976	0.3009153	-4.66	0.02325	95.7	0.0957
	p_2	0.4357385	0.4191746	-73.73	0.02790	94.6	0.1059
	p_3	0.1478475	0.1757833	166.67	0.03257	81.4	0.0845
	p_4	0.1144165	0.1041269	-61.86	0.01955	90.1	0.0674
<i>DISTANCE</i>	p_1	0.3019976	0.3132913	44.93	0.02754	93.4	0.1017
	p_2	0.4357385	0.4447191	34.79	0.02732	96.7	0.1110
	p_3	0.1478475	0.1475596	-1.25	0.02296	92.0	0.0791
	p_4	0.1144165	0.0944300	-91.95	0.02798	73.2	0.0644

Table 2: Results for the second version of the set of thresholds.

Rounding	Parameter	<i>TV</i>	<i>AE</i>	<i>SB</i>	<i>RMSE</i>	<i>CR</i>	<i>AW</i>
<i>SIMPLE</i>	p_1	0.3019976	0.2778029	-102.98	0.03223	84.1	0.0929
	p_2	0.1091671	0.1410149	210.21	0.03431	74.4	0.0784
	p_3	0.2270266	0.2431438	79.64	0.02586	94.0	0.0932
	p_4	0.3618717	0.3389044	-103.91	0.03326	86.0	0.1001
<i>DISTANCE</i>	p_1	0.3019976	0.3004218	-2.14	0.02392	95.4	0.0990
	p_2	0.1091671	0.1077877	-7.54	0.01834	91.4	0.0665
	p_3	0.2270266	0.2370702	39.75	0.02718	91.8	0.0940
	p_4	0.3618717	0.3547203	-31.40	0.02674	93.9	0.1032