Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and

Effective Tool for Handling Missing Data

Teresa A. Myers

George Mason University

Center for Climate Change Communication

Email: TeresaAnnMyers@gmail.com

Abstract

Missing data are a ubiquitous problem in quantitative communication research, yet the missing data handling practices found in most published work in communication leave much room for improvement. In this paper, problems with current practices are discussed and suggestions for improvement are offered. Finally, hot deck imputation is suggested as a practical solution to many missing data problems. A computational tool for SPSS is presented which will enable communication researchers to easily implement hot deck imputation in their own analyses.

Goodbye, Listwise Deletion: Presenting an Easy and Effective Tool for Handling Missing

Data in Communication Research

Journal articles and book chapters in fields such as sociology (Little & Rubin, 1989), political science (King, 2001), psychology (Roth, 1994), education (Peugh & Enders, 2004) and our own, communication (Harel, Zimmerman, & Dekhtyar, 2008), bemoan the lack of sophisticated practice in the handling of missing data. The common thread throughout all of these works is the impunity with which we as social science researchers continue to ignore best-practices in the arena of handling missing data. The fault, however, is not entirely on us as researchers, for with rare or no penalties for inaction, there is little impetus for change. My purpose in this paper is to raise awareness about the problems of the status quo, while simultaneously providing a user-friendly tool that quantitative communication researchers can easily implement in their data analysis strategies.

Current Practices of Communication Scholars

While we as communication researchers may admit that missing data are less than ideal, we have not spent much time as a field implementing effective strategies for addressing the problem. According to a recent content analysis of several prominent publications in the field of communication, only 22% of quantitative articles even mentioned how they handled their missing data (Harel, Zimmerman, & Dekhtyar, 2008). Given the ubiquity of missing data and the fact that each researcher must make a decision to handle the missing data in some way (even if it is choosing to use the default of listwise deletion), this absence of even a mention of procedures used for missing data

seems to indicate that the technique that a researcher implements is not currently considered to be of much importance to authors, reviewers, and editors. A tacit understanding that missing data is a trivial nuisance seems to be the rule. I argue in this paper that this unspoken assumption no longer suffices for communication research.

Based on Harel, Zimmerman, and Dekhtyar's (2008) content analysis, it seems that the de-facto manner by which most of us choose to deal with missing data is listwise deletion, meaning simply discarding any case which is missing a measurement on the variable(s) that we are interested in (also known as casewise deletion). For example, in a regression analysis predicting attention to news from the three independent variables of sex, education, and income, the majority of us would use listwise deletion to discard any case which was missing on any of the four included variables. According to Harel, Zimmerman, & Dekhtyar, 75% of those articles which mentioned the handling of missing data chose to use listwise deletion (comprising 17% of *all* quantitative articles in the content analysis, even those which mention no approach to handling missing data). A minority of communication scholars implemented some other strategy, including pair-wise deletion (1% of all quantitative articles included), mean imputation (1%), full information maximum likelihood (2%), and multiple imputation (2%).

Listwise deletion is advantageous in that it is easy to implement and is the default in many statistical packages, including SPSS. However, its ease of implementation is offset by the disadvantages accrued when deleting cases due to missing data. In the words of Harel, Zimmerman, & Dekhtyar (2008) listwise deletion is "a method that is known to be one of the worst available" (p. 351). If we make the assumption that all quantitative

articles in the aforementioned content analysis which made no mention of how they handled missing data did in fact utilize listwise deletion (an assumption which is not untenable, given that it is the default in many statistical packages), then a staggering 94% of these published communication articles used this worst possible of all methods.

Problems with the Status Quo of Handling Missing Data in Communication Research

Problems Caused by Oft-Used Methods of Missing Data Handling

In the provocatively titled *Listwise Deletion is Evil*, the problems with listwise deletion are enumerated, including that it reduces the effective sample size and introduces bias into estimates (King, Honaker, Joseph, & Scheve, 1998). In order to more completely elaborate on the problems that can be caused by listwise deletion and other such easily implemented missing data handling techniques, it is necessary to consider the various mechanisms that might produce missing data. Data can be absent for a variety of causes and the reason(s) that data are missing influence the appropriateness of strategies used to address the problem (Little & Rubin, 1989). In order of increasing seriousness to the accuracy of estimation, missing data can take one of three forms: Missing Completely at Random, Missing at Random, and Missing Not at Random. These labels are not intuitively meaningful, so it is helpful to flesh out their meanings prior to addressing the appropriateness of various missing data handling procedures under each of these patterns of missing data (See Figure 1).

Figure 1 About Here

Missing Completely at Random (MCAR). Data are considered missing completely at random when the probability of whether or not an individual is missing a value on a given measurement is unpredictable. That is, there is no systematic underlying process (except for random variation) as to why individuals are missing for a given measurement. It may be that a page of the questionnaire was accidently dropped for one participant, or that some individuals inadvertently skipped a question, or that other individuals were momentarily distracted. Data would be MCAR if (in a perfect world) we could measure all possible reasons why we might suspect individuals might choose to skip a given question and then upon testing these explanations for missingness, we find that there is no relationship between these reasons and the pattern of missingness observed. For example, if there was no way to predict whether or not someone was missing on attention to news, then attention to news would be MCAR.

Missing at Random (MAR). The second pattern is data missing at random. Data are considered MAR if they are missing because of some potentially observable, *non-random*, systematic process. The title Missing at *Random* may be a bit of an intuitive trap, however, the pattern is not difficult to understand in spite of this misnomer. Essentially, data are MAR if the probability of missingness for some variable (Y) is predictable based on the value of another variable or set of variables (X). Thus, if we were able to measure all potential X's, data would be MAR if we could predict the probability that an individual with given characteristics would be missing on Y with this set of X's. So, for example, if people who had low education were more likely to be missing on attention to news, then attention to news would be MAR.

Missing Not at Random (MNAR). Data are considered missing not at random if they are missing due to the value of the variable being considered. That is, if we are considering the pattern of missing variables on variable *Y*, it would be MNAR if individuals choose not to respond because of their true value of *Y*. A classic example is income. Income may often be MNAR because individuals who make an extremely high or low income might choose not to report the value of their income. Thus, the pattern of missingness of the income variable is dependent upon the value of an individual's income and is MNAR. Considering our example of attention to news, if people who rarely attended to news were more likely to decline to answer a question about attention to news, then attention to news would be MNAR.

Listwise Deletion Problems. The extent to which listwise deletion will cause problems in data analysis is dependent on the pattern of missingness within the data (whether it is MCAR, MAR, or MNAR). Of course, in practice we are never able to know with certainty which pattern accurately describes the pattern of missingness in the data that we possess, so we must make assumptions along the way. If the assumption of MCAR (the least serious pattern of missing data) holds, listwise deletion can still produce problems. Under MCAR, listwise deletion causes a loss of power, so that the ability to detect an existing relationship diminishes (or, more accurately, the probability of rejecting a false null hypothesis decreases). King et al. (1998) explain the problems of listwise deletion in a typical multivariate analysis, even when the best case MCAR assumption holds true (which is rarely warranted):

[On average] the point estimate… is about a standard error farther away from the truth because of listwise deletion… In some articles [it] will be too high, and in others too low, but "a standard error farther from the truth" gives us a sense of how much farther off our estimates are on average, given MCAR. This is a remarkable amount of error, as it is half of the distance from no effect to what we often refer to as a "statistically significant" coefficient (i.e., two standard errors from zero). (p. 6)

The problem is even more serious when MCAR does not hold, which is true in most instances. So, when the probability that a value will be missing is predictable, for example when people very low in attention to news decline to answer a news attention inquiry or if individuals with less education are more likely to skip answering a question about their support of a particular media policy, then the use of listwise deletion can introduce severe bias into the analysis, including altering the sign and magnitude of estimates (Anderson, Basilevsky, & Hum, 1983).

Pairwise Deletion Problems. Like listwise deletion, lesser-used methods such as pairwise deletion and mean substitution also can hinder the conclusions reached by communication researchers. Pairwise deletion discards cases on an analysis by analysis basis, and only when the estimate "requires" that variable. Thus, in a multiple regression predicting attention to news from the three independent variables of sex, education, and income, pairwise deletion would calculate the point estimate for sex discarding only cases missing on sex and attention to news (and not those missing on the other variables of education and income). In practice, this means that different participants are included

in the estimation of each separate regression coefficient. This can result in biased estimates and, at times, such a practice may lead to mathematically inconsistent results (Kim & Curry, 1977).

Mean Substitution Problems. Mean substitution involves imputing the mean of a variable in the place of any case which is missing a value for that same variable. So, for example, if a case was missing a value for education in our regression analysis predicting attention to news, the researcher would simply place the mean value of education in the place of the missing value. This method of mean substitution would allow the researcher to include that participant in all final analyses. Mean substitution has the advantage of returning a complete data set, so estimates are based off of the same cases included in each analysis. However, mean substitution also artificially deflates the variation of a variable. Furthermore, as mean substitution is replacing all missing values with "average" scores, such a technique for handling missing data has the potential to change the value of estimates.

Thus, listwise deletion, pairwise deletion, and mean substitution, while having the advantage of being easy to implement, involve unattractive concessions in statistical power and bias. Communication researchers would be advised to avoid these methods of handling missing data.

More Statistically Appropriate Methods

Although communication researchers may be aware of the problems that arise when utilizing listwise deletion, pairwise deletion, or mean substitution, the trade-offs of executing a more statistically appropriate method may hinder us from attempting to do

so. Primarily, optimal methods for handling missing data are unfamiliar to most communication researchers. Furthermore, with the exception of a few specialized data analysis packages, most of the more appropriate methods for handling missing data are inaccessible and often quite difficult to implement. I will introduce several methods below which are preferable to listwise deletion. Several are still quite computationally intensive. However, I end by suggesting a user-friendly method which can be incorporated easily and efficiently into the typical communication researcher's toolbox.

Three methods which are often recommended for handling missing data are maximum likelihood, expectation maximization, and multiple imputation (although this is not an exhaustive list of recommended methods). These methods are sometimes called model-based strategies of dealing with missing data and are not primarily concerned with replacing the missing data but rather focus on obtaining accurate estimates of parameters.

Maximum Likelihood. Maximum likelihood [ML] procedures model the missing data based on the data available. Essentially, ML procedures consider the available data as a representative sample of some distribution (see DeSarbo, Green, & Carroll, 1986). Parameters are then estimated that maximize the chance of observing the observed data. Basically, ML attempts to create models that optimize the probability of finding the relationships observed in the data (see Allison, 2002, p. 13).

Expectation Maximization. Expectation maximization [EM] is quite similar to ML procedures of handling the missing data, although the process is iterative. In EM, the first step estimates the missing data using the observed data and the first estimates of the model parameters. In the second step, these data are incorporated and parameters are

estimated incorporating the formerly missing data. This process is continued iteratively until the change in parameter estimates is negligible. The exact specifics of the computational process are complex, however, interested readers are directed to Bilmes (1998) and Enders (2001).

Multiple Imputation. Multiple imputation replaces each missing value with a set of imputed values. Essentially, through some imputation method (like hot deck), multiple complete datasets are constructed. Analyses are then repeatedly run (typically 2-5 times) and the parameter estimates are averaged across these discrete analyses (for details see Rubin, 1987; Schafer, 1999).

These more sophisticated methods (ML, EM, and MI) of dealing with missing data are available in specialized statistical packages such as EMCOV, NORM, SAS, Amelia, SPLUS, LISREL, and Mplus (Schafer & Graham, 2002). However, these approaches are not easily implemented or available in programs most commonly used by communication researchers, such as SPSS. This barrier to use inhibits many main-stream communication researchers from implementing these strategies in data analysis.

Hot Deck Imputation. Hope for improvement, however, is not lost. Roth's (1994) analysis of various strategies of handling missing data suggests that hot deck imputation is a strategy which can be both valid (under most conditions) and simultaneously easy to use (see Figure 2). Hot deck imputation involves replacing a missing value with the value of a similar "donor" in the dataset that matches the "donee" in researcher-determined categories (see Andrige & Little, 2010 and Sande, 1983 for a more thorough overview of the hot-deck imputation method; see also Hawthorne & Elliott, 2005 and

Roth, 1994 for an overview of handling missing data and comparison of methods). A hot

deck procedure will first sort the rows (i.e. respondents) of a data file within a set of

variables, called the "deck" (also known as adjustment cells, see Andrige & Little, 2010

and Brick & Kalton, 1996). These "deck variables" are chosen by the researcher and

should typically include (a) little to no missing data, (b) discrete values (rather than

continuous variables, although continuous variables can be categorized in order to use in

the deck), and (c) related to the variable being imputed or predictive of non-response –

but not of substantial theoretical interest to the research questions being addressed.

Figure 2 About Here

Respondents with complete data who match on all deck variables to a respondent

who is missing on the variable in question (the "donee") are eligible to donate their score

to that respondent. After sorting respondents into these decks, all respondents within a

given deck are randomly sorted and any respondent missing on a given variable is then

assigned the value of respondent nearest to him or her in this randomly permuted data file

who is not missing data. This method has the effect of assigning a response to

nonresponses by randomly sampling without replacement from the distribution of the

responses to that question from other respondents with the same set of values on the deck

variables as the respondent. Thus, just as fluent English speakers are able to use the

information available in a sent_nce to f_gure out what words with miss_ng l_tters are

supposed to be, hot deck imputation uses information that is available in the data to "fill in" information that is missing.

Like many missing data procedures that are more appropriate than listwise and pairwise deletion, hot deck imputation is not available in programs communication researchers use widely and often. In order to facilitate its greater adoption, a computational aide in the form of an SPSS macro is presented in the Appendix. The HOTDECK macro creates a new command for SPSS users that will allow them to easily perform hot deck imputation on missing data. After running the set of commands in SPSS as an SPSS syntax file, the researcher can simply employ the syntax command "hotdeck" in the following structure: HOTDECK y = *name of the original variables which are missing values*/deck = *the set of "deck" variables*.

Figure 3 About Here

To illustrate this process, consider the data in Figure 3. The table on the left shows that participants "A" and "K" are missing on the variable *attention to news* (labeled NewsAttn). To perform a hot deck imputation on NewsAttn, first the complete text of the macro in the Appendix would be entered into SPSS syntax without alterations and run. Next, variables defining the decks would be chosen. In this example, the variables *sex, education,* and *income* were chosen to define the decks because they have no missing data and are related (but, hypothetically, not of substantive interest for the research questions) to the variable *attention to news*. Finally, the following syntax would be

employed: HOTDECK y=NewsAttn/deck = sex education income. This will create a new variable, *NewsAttnHD*, with missing values imputed and with variable and value labels from *NewsAttn* copied to *NewsAttnHD*. This process would be completed for each variable in the analysis that has missing data and then analysis would proceed as normal. Several variables can be listed in the HOTDECK command, and hot deck imputation will be conducted simultaneously for all in the list (e.g. the syntax HOTDECK y = NewsAttn NewsExposure Attitude1/deck = sex education income would produce the new variables NewsAttnHD, NewsExposureHD, and Attitude1HD, with values imputed for previously missing data).

The table on the right in Figure 3 might help elucidate what happens internally in the macro. First, rows in the data file are automatically sorted in ascending order by sex (as can be seen – participants H thru N, who are coded "0" on sex are prior in the data set to participants G thru B who are coded "1"). Second, rows in the data file are sorted in ascending order by education *within* sex (so that participants H thru Q, who are coded "0" on sex and "1" on education are prior to participants L and V, who are coded "0" on sex and "3" on education, and so forth). Third, rows in the data file are sorted in ascending order by attention to news *within* matches on both sex <u>and</u> education (so that participants H and K, who are coded "0" on sex, "1" on education, and "1" on income, are prior to participants F and Q, who are coded "0" on sex, "1" on education, and "4" and "5", respectively on income). The macro then automatically generates a random number and sorts participants in ascending order by this random number within those participants who match on the deck variables of sex, education, and income (thus, H, coded "0" on

sex, "1" on education, "1" on income, with random number ".89" is placed prior to K, coded "0" on sex, "1" on education, "1" on income, and random number ".95"). After this matching and sorting process, the macro assigns any case missing on attention to news the value of the case immediately *under* the missing case in the data file. As a result of this random sorting, the donor case is essentially randomly chosen from all cases with complete data in that case within the deck.

In some circumstances, this donor case will also be missing or will not belong to the same deck (if the missing case is the last case in the deck after sorting, for instance) as the case being imputed. When this occurs, then the macro will use the case above the missing case as the donor (and case K matches with case H, thus K receives H's value). If neither of those rows match (or if they are also missing data), then the macro will search two rows below for a donor, and then two rows above if a donor is still not found. If a donor case is not found after the completion of these iterations, the missing case is left as missing. After executing the HOTDECK macro, the user should check that all missing data have been replaced. In rare circumstances, some cases may still be missing data. If this occurs, delete the imputed variables and reexecute the macro until all missing data are successfully replaced (it may also be necessary to examine the choice of deck variables – see discussion below).

Strengths and Weaknesses of Hot Deck Imputation

The use of the hot deck imputation does have several limitations. The first is that unique cases—cases that are dissimilar to the all others in the data set on the combination of sorting variables so that no "deck match" can be found—produce a problem. For

example, in Figure 3, if participant "L" was missing on attention to news, no other participant in this simulated data file could be found who matches participant L on the variables sex, education, and income. Thus, there would be no "donor" available. This situation occurs more often in small data sets, when many sorting variables are used, when decks are defined by continuous variables, or when decks are defined by variables with many unique values. It is optimal to balance the size of the file with the number of sorting variables. A larger file can support the use of more sorting variables than a smaller file. Another problem noted by Siddique and Belin (2008) note that single hot deck procedures "fail to account for the uncertainty due to the fact that the analyst does not know the values that might have been observed" (p. 84). Multiple imputation procedures are thought to better handle this uncertainty.

Although imperfect, the hot deck method of handling missing data offers several advantages over listwise and casewise deletion. Primarily, hot deck procedures allow for retention of the complete sample of individuals, avoiding the loss of incomplete cases and the subsequent declines in statistical power that are incurred as a result. Siddique and Belin (2008) argue that the benefits of hot deck imputation include that: "(1) imputations tend to be realistic since they are based on values observed elsewhere; (2) imputations will not be outside the range of possible values (as might happen with multiple imputations, see He, 2010); and (3) it is not necessary to define an explicit model for the distribution of the missing values" (p. 84; see also Andrige & Little, 2010; Roth, Switzer, & Switzer, 1999). In their comparison of various techniques of handling missing data, Hawthorne and Elliot (2005) found hot deck imputation to be over *80 times* more

effective than list-wise deletion and that hot deck imputation also outperformed pairwise deletion and mean substitution. Furthermore, users of hotdeck imputation are in good company, as many prominent large-scale surveys implement hot deck procedures to deal with missing data, including the U.S. and British Censuses, the Current Population Survey, the Canadian Census of Construction, the U.S. Annual Survey of Manufacturers and the U.S. National Medical Care Utilization and Expenditure Survey (see Roth, Switzer, & Switzer, 1999). Hot deck imputation is recommended by Roth (1994) for all missing data scenarios, except those where the data are MNAR and constitute greater than 10% of the sample (in which case ML, MI, and EM techniques are recommended; see Figure 2). Finally, the relative simplicity of the hot deck technique in comparison to model based techniques makes it an attractive alternative to listwise deletion and has the potential to facilitate wide use and application.

Hotdeck imputation is a statistically valid approach for many missing data problems. Given the benefits of hot deck imputation and the ease with which it can now be incorporated into one's analysis plan using the SPSS tool introduced here, it is my hope that many communication researchers will soon consider the use of hotdeck imputation over demonstrably inferior approaches when they encounter missing data.

References

Allison, P. D. (2001). *Missing data.* Los Angeles: Sage publications.

Anderson, A. B., Basilevsky, A., & Hum, D. (1983). Missing data: A review of the

    literature. In P. Rossi, J. Wright, & A. Anderson (Eds.), *Handbook of survey*

    research (pp. 415-494). San Diego: Academic Press.

Andridge, R.R. & Little, R.J.A. (2010). A review of hot deck imputation for survey non-

    response. *International Statistical Review, 78*(1), 40-64.

Bilmes, J. (1998) A gentle tutorial of the EM algorithm and its applications to parameter

    estimation for Gaussian mixture and hidden Markov models. *International*

    Computer Science Institute Technical Report, International Computer Science

    Institute. Available at

    http://ssli.ee.washington.edu/people/bilmes/mypapers/em.pdf

Brick, J.M. & Kalton, G. (1996). Handling missing data in survey research. *Statistical*

    Methods in Medical Research, 5(3), 215-238.

DeSarbo, W.S., Green, P.E., & Carroll, J.D. (1986). Missing data in product-concept

    testing. *Decision Sciences, 17,* 163-185.

Enders, C. K. (2001). A primer of maximum likelihood algorithms available for use with

    missing data. *Structural Equation Modeling , 8*, 128-141.

Harel, O., Zimmerman, R., & Dekhtyar, O. (2008). Approaches to the handling of

missing data in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The SAGE Sourcebook of Advanced Data Analysis Methods for Communication Research* (pp. 349-371). Los Angeles: Sage Publications.

Hawthorne, G. & Elliott, P. (2005). Imputing cross-sectional missing data: Comparison of common techniques. *Australian and New Zealand Journal of Psychiatry, 39*(7), 583-590.

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research , 6*, 215-240.

King, G. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review , 95*, 49-69.

King, G., Honaker, J., Joseph, A., & Scheve, K. (1998). Listwise deletion is evil: What to do about missing data in political science.

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

Peugh, J. L., & Enders, C. K. (2004). Missing data in education research: A review of reporting practices and suggestions for improvement. *Review of Educational Research , 74* (4), 525-556.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology , 47* (3), 537-560.

Roth, P.L., Switzer, F.S., & Switzer, D.M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods, 2*(3), 211-232.

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

Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys, 3,* 339-349.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods , 7* (2), 147-177.

Siddique, J. & Belin, T.R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine, 27*, 83-102.

Footnotes

[1] Reilly (1992, p. 308) notes that the precision gain using hotdeck imputation is maximized when the "auxiliarly covariate(s) ("deck" variables) are highly informative about the missing X…for non-informative Z ("deck" variables), there is no gain in precision, but neither is there any penalty" (see also Andrige & Little, 2010, p. 43).

[2] As defined by effectiveness in estimating the true (known) t-value from a data set with randomly generated missing values (see Hawthorne & Elliott, 2005, p. 588).

Appendix

Execute the command set below in an SPSS syntax window *exactly as is*. Do not modify this code at all. Once executed, the HOTDECK command can be given in a new syntax window, as documented in this article. The syntax structure is

HOTDECK y = *variables with missing data*/deck = *variables defining the decks*.

An electronic version of this code can be obtained by emailing the author at TeresaAnnMyers@gmail.com. The file is also available on the web. To find its current location, search for "SPSS Hot deck macro" using your favorite web browser.

```
DEFINE HOTDECK (y = !charend ('/')/deck = !charend ("/")).
Output New name =  hotdeckextra.
!do !s !in (!y).
compute randnum = uniform(1).
sort cases by !deck randnum.
compute sortclg1 = 1.
compute sortclg2 = 1.
compute sortcld1 = 1.
compute sortcld2 = 1.
!DO !v !in (!deck).
create sortd1v = lead(!v,1).
create sortd2v = lead(!v,2).
if (lag(!v) <> !v) sortclg1 = 0.
if (lag(!v,2) <> !v) sortclg2 = 0.
if (sortd1v <> !v) sortcld1 = 0.
if (sortd2v <> !v) sortcld2 = 0.
!DOEND.
!let !newname = !CONCAT (!s, HD).
compute newvar = !s.
apply dictionary from * /source variables = !s /target
variables = newvar.
execute.
Create yLead = Lead(!s,1).
Create yLead2 = Lead (!s,2).
DO If (Missing(newvar)).
+    DO IF ((sortclg1 = 1) AND Not Missing(lag(!s))).
+         Compute newvar = Lag(!s).
+    ELSE IF ((sortcld1 = 1) AND Not Missing (yLead)).
+         Compute newvar = yLead.
+    ELSE IF ((sortclg2 = 1) AND Not Missing(Lag(!s,2))).
+         Compute newvar = Lag(!s,2).
```

```
+     ELSE IF ((sortcld2 = 1) AND Not Missing(yLead2)).
+         Compute newvar = yLead2.
+     END IF.
End If.
Match Files/File = */drop yLead ylead2 sortd1v sortd2v
sortclg1 sortclg2 sortcld1 sortcld2 randnum.
execute.
rename variables (newvar = !newname).
!doend.
output close name = hotdeckextra.
!ENDDEFINE.
```

Figure 1
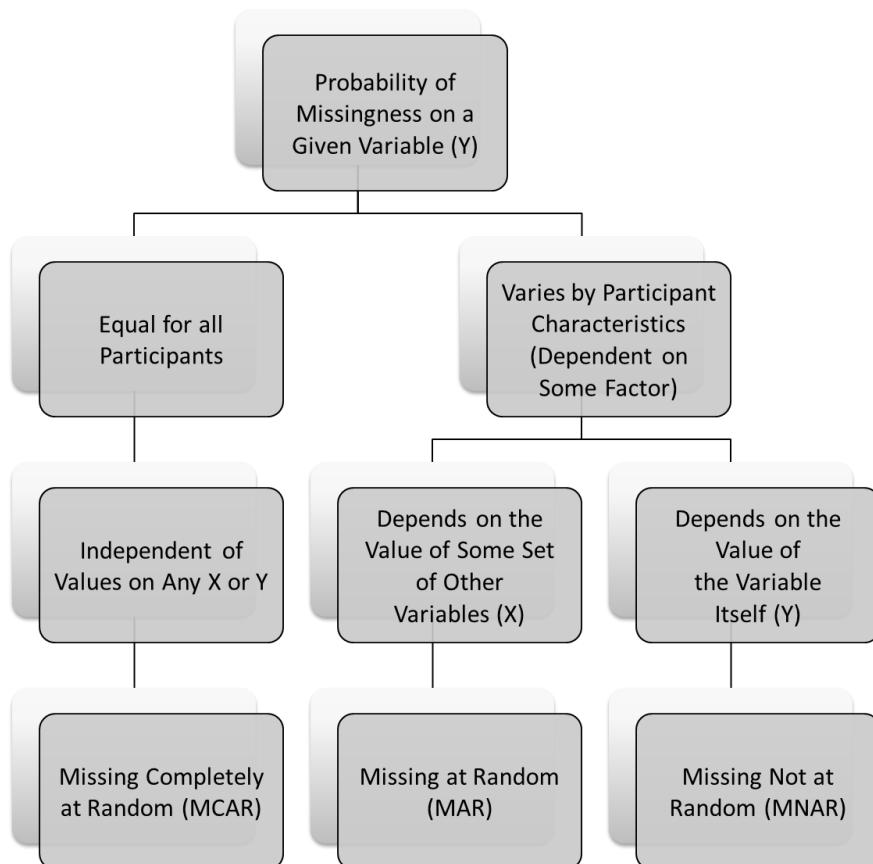
Figure 2

Percentage of Missing Data	Pattern of Missing Data		
	Missing Completely at Random	Missing at Random	Missing Not at Random
1-5%			
6-10%	Hot Deck Recommended		
11-15%			Utilize Maximum Likelihood, Expectation Maximization, or Multiple Imputation
16-20%			

Figure 3

Original Data File				
Participant	Sex	Education	Income	NewsAttn
A	1	3	3	
B	1	5	5	1
C	1	2	2	5
D	1	5	5	2
E	1	4	4	5
F	0	1	4	5
G	1	1	3	2
H	0	1	1	1
I	1	3	3	3
J	1	3	3	2
K	0	1	1	
L	0	3	4	4
M	1	5	2	1
N	0	5	2	1
O	1	2	3	2
P	1	4	1	4
Q	0	1	5	5
R	1	5	4	4
S	0	4	3	1
T	0	4	4	1
U	0	4	2	3
V	0	3	5	1
W	1	2	1	2
X	0	4	2	4
Y	1	2	4	3
Z	1	4	5	2

Data File Randomly Sorted Within Decks						
Participant	Sex	Education	Income	NewsAttn	RandomNumber	NewsAttnHD
H	0	1	1	1	0.89	1
K	0	1	1		0.95	1
F	0	1	4	5	0.75	5
Q	0	1	5	5	0.12	5
L	0	3	4	4	0.79	4
V	0	3	5	1	0.10	1
X	0	4	2	4	0.02	4
U	0	4	2	3	0.61	3
S	0	4	3	1	0.61	1
T	0	4	4	1	0.66	1
N	0	5	2	1	0.37	1
G	1	1	3	2	0.99	2
W	1	2	1	2	0.49	2
C	1	2	2	5	0.86	5
O	1	2	3	2	0.37	2
Y	1	2	4	3	0.87	3
I	1	3	3	3	0.36	3
J	1	3	3	2	0.79	2
A	1	3	3		0.81	2
P	1	4	1	4	0.48	4
E	1	4	4	5	0.56	5
Z	1	4	5	2	0.95	2
M	1	5	2	1	0.11	1
R	1	5	4	4	0.83	4
D	1	5	5	2	0.20	2
B	1	5	5	1	0.57	1

Note: Data were randomly generated for demonstration purposes only.