ARTICLE
Methods for Handling Missing Data in the Behavioral Neurosciences: Don't Throw the Baby Rat out with the Bath Water

Leah H. Rubin,[1,2] Katie Witkiewitz,[1] Justin St. Andre,[1] and Steve Reilly[1]
[1] Department of Psychology, University of Illinois at Chicago, Chicago, IL 60612; [2] Department of Psychiatry, University of Illinois at Chicago, Center for Cognitive Medicine, Chicago, IL 60607.

Missing data are a major problem in the behavioral neurosciences, particularly when data collection is costly. Often researchers exclude cases with missing data, which can result in biased estimates and reduced power. Trying to avoid the deletion of a case because of a missing data point can be conducted, but implementing a naïve missing data method can result in distorted estimates and incorrect conclusions. New approaches for handling missing data have been developed but these techniques are not typically included in undergraduate research methods texts. The topic of missing data techniques would be useful for teaching research methods and for helping students with their research projects. This paper aimed to illustrate that estimating missing data is often more efficacious than complete case analysis, otherwise known as listwise deletion. Longitudinal data was obtained from an experiment examining the effects of an anorectic drug on food consumption in a small sample (*n*=17) of rats. The complete dataset was degraded by removing a percentage of datapoints (1-5%, 10%). Four missing data techniques: listwise deletion, mean substitution, regression, and expectation-maximization (EM) were applied to all six datasets to ensure that each approach was applied to the same missing data points. *P*-values, effect sizes, and Bayes factors were computed. Results demonstrated listwise deletion was the least effective method. EM and regression imputation were the preferred methods when more than 5% of the data were missing. Based on these findings it is recommended that researchers avoid using listwise deletion and consider alternative missing data techniques.

Key words: missing data, imputation, expectation maximization, listwise deletion, mean substitution, regression

Statistical analyses of behavioral data are often limited by multiple missing values attributable to a variety of sources (e.g., attrition, illness). The tendency for researchers is to exclude any cases with missing values and/or use qualitative methods to draw conclusions about their data. Unfortunately, many of the statistical methods used to analyze behavioral data, including repeated measures and multivariate analysis of variance (ANOVA) as well as multiple regression techniques, require each sample to have a complete array of values. Studies with small samples and missing data may result in too few cases to run analyses if cases with missing values are excluded. Moreover, incomplete data can affect the conclusions of studies in a variety of ways. For example, incomplete data results in reduced statistical power for conducting hypothesis tests as well as decreased precision of estimation because of the reduced sample size. Consequently, acceptable methods for incorporating missing data are needed to increase statistical power and to allow for accurate estimation of statistical effects. The purpose of this paper was to alert researchers to the problem of missing data, to illustrate that in certain situations estimating missing data leads to more accurate estimates than listwise deletion, and to help guide researchers in selecting one missing data method over another. The topic of missing data is an important one and a discussion of the techniques would be useful for teaching research methods and for helping students with their research projects.

Current Practices

Behavioral neuroscientists are able to exercise a large amount of control during experimentation relative to other disciplines resulting in less frequent missing data and, when occurring, to a lesser degree than in other fields. However when missing data does occur, a common practice among some but not all (examples: Tkacs et al., 1997; Evenden, 1999; Gonzales and Weiss, 1998; Sokoloff et al., 2000; Sokoloff and Blumberg, 2001) behavioral neuroscientists is to exclude subjects with missing data before conducting analyses. As such, the problem of missing data is avoided at the expense of power. To better handle missing data without losing power other methods of data management are available. In the remainder of this article we discuss the types of missing data, four commonly used methods for handling missing data, and demonstrate the utility of each method using a longitudinal study on feeding consumption.

Types of Missing Data

Three types of missing data have been described (Rubin, 1977; Little and Rubin, 1989). First, missing completely at random (MCAR), which implies no systematic reasons for missingness, is the standard assumption in most applications. The pattern of MCAR occurs when missing values on a variable (x) are not dependent on any values on any measured variables including the variable (x) in the dataset. Simply, the observed values are a random subset of the theoretically complete dataset. To illustrate,

suppose a researcher is conducting a study on the effects of a drug on food consumption (measured by the number of bar presses dispensing pellets) in a sample of rats. An MCAR pattern would result in this study if for some rats the number of bar presses was unobtainable due to random factors such as the pellet dispenser malfunctioning. Assuming these random factors are unrelated to bar presses, the observed data from the remainder of the rats would constitute a random subset of the theoretically complete dataset. Assuming MCAR is convenient for many researchers but it can introduce biases if the data are not truly MCAR.

Missing at random (MAR) implies that the data are not MCAR, and that some information as to why the data are missing is known and is examinable by other collected variables. Specifically, the pattern of MAR occurs when missing values of a variable (x) is related to other measured variables but the missing data is not a product of the variable (x) itself. Continuing with food consumption example, suppose that after collecting and analyzing the data the researcher finds that there is a relationship between the number of bar presses for food (x) and body weight (y). Let us assume that the obese rats were more likely than non-obese rats to press the bar for food. In this example, the probability of missing pellet consumption (x) is related to a second measured variable (i.e., weight; y). With an MAR pattern, the variables with missing data can be predicted from other acquired measures (i.e., regression equation). Missing data can be ignored if either the assumption of MCAR or MAR is met.

If the pattern of missingness is in some way related to the outcome variables, then the data are said to be Non-ignorable missing (NIM). Unlike MAR, NIM is not predictable from other variables but rather is only explainable by the variable on which missing data exists. Specifically, a NIM pattern results when missing values on a variable (x) is related to the values on (x). Returning to the food consumption example, suppose that after analyzing the data the researcher found that the missingness on food consumption did not relate to the other measured variables. While the lack of bar pressing to obtain the pellets might be misconstrued as missing data, the missing data may have been a function of the animal's lack of motivation to press the bar for pellets. In the above example, missing pellet consumption would be considered NIM because the missing data is only explainable by the variable, pellet consumption.

Methods for Dealing with Missing Data

Missing data has traditionally been handled by eliminating any cases with missing measurements, a technique called listwise deletion. Listwise deletion tends to be the default procedure in statistical packages such as SPSS and SAS. This procedure excludes cases with missing scores on any variable or variables used in an analysis. For instance, an equipment malfunction for a single rat on a single day of a 30-day study would result in that rat's data being excluded from all longitudinal data analyses. This method is problematic because data derived from collection procedures that were time consuming or costly are not incorporated into the analysis. When data are NIM listwise deletion can introduce systematic biases defined by the patterns of missingness; and if data are MAR or MCAR, then listwise deletion will result in a reduced sample size and less power to detect statistical effects (Allison, 2002). When data are missing and not NIM, then it is important to include them if possible (Schafer and Graham, 2002).

Several techniques for dealing with missing data have been proposed (Schafer and Graham, 2002). The two most accepted and recommended methods for handling missing data are maximum likelihood (ML; Little and Rubin, 1987) and multiple imputation (MI; Rubin, 1977; MI; Schafer, 1999). The results from MI and ML are very similar under most conditions and neither has been found to be superior (Collins et al., 2001). In ML, parameter values with the highest possible probability are assigned using the probability density of the realized data, called a likelihood function. Using this likelihood function the ML procedure provides parameter estimates based on all available data, including the incomplete cases. However, simulation studies show that ML is an inadequate estimation technique for some small sample problems and results in biased estimates (Little and Rubin, 1989). For large samples ML is a preferred method for dealing with missing data (Schafer and Graham, 2002). In MI, statistical models are built to "fill in" (or impute) missing data. MI uses all available data to estimate a distribution of possible values for each missing data point after which random error can be estimated by combining the possible distributions of each data point for a pre-specified number of multiply imputed data sets. The decision to use MI or ML can be based on the individual researcher's specific questions and preferred data analytic techniques. Unfortunately, both MI and ML methods are less accessible to many researchers because they require either large sample sizes (ML) or special statistical software (MI). Since we have a small sample and these are large sample techniques requiring more sophisticated software programs, we do not address ML and MI in this paper.

The most commonly practiced approaches are mean substitution and regression based methods—both single imputation techniques. Mean substitution replaces missing values on a variable with the mean value of the observed values. The imputed missing values are contingent upon one and only one variable – the between subjects mean for that variable based on the available data. Mean substitution preserves the mean of a variables distribution; however, mean substitution typically distorts other characteristics of a variables distribution (i.e., variance, median; Little and Rubin, 1989). For example, mean substitution restricts the variability of a variable and alters the underlying distribution to be more peaked at the mean (Allison, 2002). Due to these distributional problems, statisticians often suggest ignoring missing values rather than imputing values by mean substitution (Little and Rubin, 1989).

Regression based methods are a more sophisticated technique than mean substitution and imputes missing data from multiple variables by replacing missing values for

a variable (criterion; Y) by using observed values for a case (predictors, $X_1 \ldots\ldots X_n$). Essentially, a regression equation is being created $\hat{Y}_i = b_{0i} + x_{1i}b_{1i}$ + error—where $\hat{Y}_i$ is the predicted missing data value for person *i*, b_{0i} is the intercept (constant) for person *i*, *x* is the score on variable *x* for person *i,* and b_{1i} is the slope coefficient for subject *i.*

Regression techniques are considered conditional approaches because missing values are conditional upon the predictors that are incorporated into the regression equation (Little and Rubin, 1989). To illustrate, let us go back to the food consumption example where the researcher was interested in determining the effects of a drug on food consumption. Let us assume that the researcher was missing data on some of the rats weight. Let us also assume that the research collected data on the amount of water intake and exercise and that these variables were relevant and moderately correlated to weight. To generate the missing values on weight, the researcher would use the water intake and exercise variables as the predictor variables in the regression equation where weight was the predicted variable. The regression equation would be:

$weight_i = intercept_{0i} + water\ intake_{1i}b_{1i} + exercise_{2i}b_{2i}$ + error.

Despite the attraction of this method, it is still not suggested by researchers conducting correlations or analysis of covariance (ANCOVA) because the percentage of variance explained (R^2) is assumed to be perfect (1.00) and the estimates of variability are underestimated (Little and Rubin, 1989; Enders, 2001). Consequently, regression based methods can overestimate the relationships between the predictor(s) and criterion and increase the likelihood of Type I errors (Schafer and Graham, 2002). Some programs, such as SPSS, provide the option of adding random error to the equation to adjust for overestimation the relationship between the Xs and Y which may reduce the likelihood of a Type I error.

Expectation maximization (EM) and maximum likelihood (ML; described below) are both missing data methods that provide a maximum-likelihood estimate of the covariance structure given the available data. EM is a two-step process. In the Expectation step, the "expected values" for missing observations are computed using regression equations given the observed data and the missing observation is replaced by the conditional mean based on the regression equations (Dempster et al., 1977). In the Maximization step the estimates are updated to maximize the log likelihood based on the statistics from the Expectation step. This two-step process is repeated for a user-specified number of iterations. The EM algorithm has advantages in satisfying the Missing at Random assumption (MAR), because supplementary variables can be included in the initial regressions of the Expectation step (Collins et al., 2001; Enders and Peugh, 2004). Exact sample sizes needed for EM are not known and recent simulations have identified complexity in the determination of a necessary sample size for using the EM algorithm (Enders and Peugh, 2004). In the Enders and Peugh (2004) simulation the authors concluded that no single value of *n* is appropriate for EM and two separate analyses (one for model fit and one for testing significance of parameters) should be conducted. An exception to this rule is when the researcher is only interested in estimating individual model parameters. When one model parameter is tested at a time then the mean parameter value yielded accurate coverage of the confidence interval estimates across many model simulation conditions. Choi et al. (2004) conducted a simulation study on EM and demonstrated the parameter estimates were consistent with the estimates from the complete data, even with up to 50% of the data missing.

Therefore, the current study was designed to determine which of the alternative missing data techniques (mean substitution, EM, regression imputation) compared to listwise deletion would provide the most robust and consistent estimates of the complete data, after the complete data had been degraded by removing a percentage of the cases. Based on previous reviews of these missing data methods (Little and Rubin, 1987; Roth, 1994; Wothke, 1998), it was hypothesized that EM and regression imputation would provide the most accurate estimates at all levels of missing data and listwise deletion would provide the least accurate estimates. Mean substitution was expected to be inferior to EM and regression imputation, but superior to listwise deletion.

MATERIALS AND METHODS

The data reported in this paper were collected in an effort to examine the effects of an anorectic drug on feeding behaviors in Sprague-Dawley rats (*n* = 17) over a 21-day period (St. Andre and Reilly, unpublished data). The testing procedure entailed placing the animals in their assigned operant chamber at noon daily. On days four to seven animals were removed one at a time and given an i.p. saline injection (sterile NaCl) and then placed immediately back into their operant chamber. The total number of pellets consumed during these three pre-test days was used to place animals in one of two conditions: saline, anorectic drug. After day seven, animals were given daily administration of either the anorectic drug treatment or vehicle for one week. Similar to days four to seven, animals were removed from their chamber one at a time, injected i.p. and placed immediately back into their chamber. Animals were injected with either an anorectic drug or vehicle (1 ml saline). Following the week of the anorectic drug or saline injections the animals had seven days without injections.

Data Analysis

For the purpose of the missing value analysis illustration, we incorporated data from the number of pellets consumed during days 7-15 of the study and compared rats injected with saline or an anorectic drug during these days. All animals provided complete data (no missing values) on all days of the study. This initial dataset, with complete data, was degraded into six new datasets using a random number generator. The six datasets were designed to have one of six levels of missingness (1-5%, and 10%). Each of the four missing data procedures (listwise deletion, mean substitution, regression, EM) were applied to each of these six datasets. This was done so each missing data

approach was applied to the exact same missing data points at each level of missingness. SPSS missing value analysis software (MVA; version 14.0 for Windows; SPSS, Chicago, IL) was used for the regression and EM imputations. To produce the imputed values for regression and EM, we used Condition (saline, anorectic drug) and food consumption across Days 7-15.

Three methods were used to compare the missing data approaches: 1) *p*-values, 2) effect sizes, and 3) Bayes factors. The classical frequentist concept of *p*-values was selected for model comparison because *p*-values are the most frequently used summary measure in statistical hypothesis testing. *P*-values for the full data as well as the 24 generated datasets were obtained from a series of mixed design analysis of variance (ANOVA) with Condition (saline, anorectic drug) as the between-subjects factor and Day (7—15) as the within-subjects factor. To account for violations in sphericity and to better establish validity of using this particular data set, we used a Geiser-Greenhouse F-test (Greenhouse and Geisser, 1959). This test adjusts the degrees of freedom so that the F-critical value will be somewhat larger and there is less of a probability of committing a Type 1 error.

Ideally in the analysis of longitudinal data, mixed-effects regression models (MRM) also known as hierarchical linear models (HLM) is the more appropriate analysis than the traditional mixed design ANOVA. MRMs can accommodate variable spaced measurements over time; repeated measurements that are correlated to different degrees; non-constant variability; time varying and/or time invariant covariates (Raudenbush and Bryk, 2002); and missing data (Rubin, 1976; Little, 1979; Hedeker and Gibbons, 1997). However, MRMs use maximum likelihood estimation which as previously mentioned yields biased estimates when sample sizes are small (Little and Rubin, 1989). Therefore, we chose to use mixed design ANOVAs.

The traditional approach of using *p*-values for hypothesis testing has received much criticism over the years. Researchers interested in an in-depth discussion of the inadequacies and criticisms of *p*-values are referred to Cohen (1994). Therefore, we extended our data analysis in two ways. First, effect sizes were computed from the mixed design ANOVAs using eta squared. Effect sizes measure the magnitude of an effect which *p*-values do not provide.

The data was also analyzed using a Bayesian approach (Bayes theorem) to hypothesis testing (Bernardo and Smith, 1994). Bayes theorem is comprised of two components: 1) prior beliefs (hypotheses) and 2) data (weight of evidence). Essentially, this approach takes a researcher's prior beliefs and examines how the data alters these beliefs. For comparing the missing data methods, we used the data component, commonly referred to as the Bayes factor which can be expressed as,

$$B = \frac{P(H_o \mid data)/P(H_a \mid data)}{P(H_o)/P(H_a)}.$$

This formula expresses how much the data changes the odds for the null hypothesis (H_o) relative to the initial prior odds/alternative hypothesis (H_a). For an in-depth discussion of Bayes factors the readers are referred to Kass and Raftery (1995). For this paper, Bayes factors were computed based on the recent methods of Sellke et al. (2001). Sellke and colleagues derived a simple formula, where given the *p*-value researchers can obtain the minimum value of the Bayes factor that one can obtain for the null hypothesis. This Bayesian calibration can be computed by $B(p) = -e^* p^* \log(p)$ where *e* is the mathematical constant (2.71828182845; base of the natural logarithm) and the lower bound on the odds for the Bayes factor uses the *p*-value. The minimum value or lower bound of the Bayes factor can be considered the smallest amount of evidence supporting the null hypothesis based on the data. For instance, a Bayes factor of .30 would mean that the null hypothesis gets 30% as much support as the best supported hypothesis. As the Bayes factor increases so does the support for the null hypothesis, and vice versa. For our purposes, Sellke's method is sufficient and easy to compute. However, it is important to note that Sellke's formula accounts for the first component of Bayes theorem (the prior probability) by assuming that the H_o and H_a have equal prior probabilities of 0.5 (Sellke et al., 2001). Therefore, the lower bound of the Bayes factor can be considered a compromise between the frequentist and Bayesian approaches.

RESULTS

Table 1 presents the means and standard deviations for pellet consumption for rats as a function of condition for the complete dataset. To compare missing data methods, Condition (saline, anorectic drug) means and standard deviations for pellet consumption were averaged across days 7-15 for each percentage of Missing Data (1-5,10%) and missing data method. Residual means and variances were then calculated by subtracting the estimated overall means and variances from the actual overall mean and variances from the dataset for each percentage of missing data, and for each missing data method for the saline and anorectic drug conditions. The results of the residuals are presented in Table 2.

Based on previous research we expected little variability in pellet consumption for animals in the saline condition and as expected there were small differences between the missing data methods on estimates for this condition. In the anorectic drug condition, there were no differences between the estimated means across all missing data methods when 1 or 2% of the data was missing thus the residual means and variances all equal 0. Greater than 2% missing data resulted in listwise deletion producing the most inaccurate mean estimates (as compared to the total sample means).

The *p*-values (see Figure 1) and effect sizes (η^2; see Figure 2) that were computed from the mixed design ANOVAs are plotted as a function of percent of missing data and missing data technique for the main effects of Day and Condition and the Day x Condition interaction on pellet consumption. Overall, both frequentist approaches indicated that there were large differences between listwise deletion and the other methods. Listwise deletion resulted in *p*-values and effect sizes that were either much greater or less than the *p*-value and effect sizes from the complete

	Condition			
	Saline		Anorectic Drug	
Day	*M*	*SD*	*M*	*SD*
7	575.63	74.84	515.22	85.20
8	547.00	81.20	510.89	45.45
9	570.50	68.80	465.33	141.42
10	552.00	75.61	476.56	137.82
11	549.25	77.39	497.33	76.57
12	570.88	57.87	496.67	90.63
13	569.13	63.08	500.67	98.01
14	543.13	73.46	538.22	86.25
15	562.00	77.12	599.67	77.72

Table 1. Means and standard deviations for pellet consumption for rats as a function of condition for the complete dataset.

dataset. With 1 to 5% missing data, mean substitution, regression, and EM were efficient procedures according to both frequentist approaches. As the percent of missing data increased (10%), the frequentist approaches showed that EM had a slight advantage over regression and mean substitution and regression had a small advantage over mean substitution. However, 2% or more missing data combined with a small sample size appears to increase the likelihood of committing Type I and Type II errors (see Figure 1). Greater Type I error (incorrectly rejecting the null) is indicated by estimated *p*-values that are lower than the actual p-value derived from the complete data and greater Type II error (incorrectly failing to reject the null, when it should be rejected) is indicated by *p*-values that are greater than the actual *p*-value derived from complete data.

Estimated effect sizes indicate that when there was greater than 2% missing data, estimation methods result in either an over or underestimate of the percent of variance explained (see Figure 2). Specifically, listwise deletion tended to underestimate the main effect for condition where it overestimated the main effect for Day and the Condition x Day interaction. When there was greater than 2% missing data, all estimation methods resulted in a slight overestimation of the magnitude of effects of the independent variables. It is important to note that all methods but listwise deletion produced accurate estimates of Condition differences. However, all methods were less accurate for estimating the main effect of Day and the Condition x Day interaction. These differences were a direct result of the number of missing values in each cell when conducting ANOVA with the degraded data and do not reflect limitations of the missing data methods used in the current study.

Figure 3 provides the lower bounds of the Bayes factors (Bayesian approach) for the effect of Day, Condition, and the Day x Condition interaction on pellet consumption that was plotted as a function of percent of missing data and missing data technique. A Bayesian framework provided a similar pattern of results as the frequentist approaches.

	Missing Data Technique							
	Listwise Deletion		Mean Substitution		Regression		EM	
Missing Data (%)	*M*	*Var*	*M*	*Var*	*M*	*Var*	*M*	*Var*
	Saline Condition							
1	-0.60	65.41	-1.63	-377.01	-1.77	-334.24	-1.63	-350.51
2	-0.01	-25.62	-2.76	-392.02	-2.68	-376.19	-2.64	-372.80
3	-1.73	-31.11	-3.26	-515.05	-3.11	-493.54	-3.11	-482.15
4	-1.37	44.31	-1.67	56.35	-1.70	57.36	-1.43	47.10
5	-0.52	241.01	-2.00	-246.02	-1.89	-238.15	-1.84	-223.44
10	1.38	-83.77	-2.01	-597.20	-2.47	-546.39	-1.71	-585.06
	Anorectic Drug Condition							
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	-1.51	160.22	0.44	-215.53	0.44	-207.46	-0.03	-228.31
4	-3.04	238.74	-1.13	-156.90	-1.22	-147.22	-0.73	-179.00
5	0.65	150.09	3.28	-374.22	3.22	-357.33	3.19	-365.51
10	-5.04	1020.10	0.79	410.38	0.41	404.99	0.47	387.87

Table 2. Residual means and variances for pellet consumption averaged across days 7-15 for each percent Missing data and missing data method. Residual means and variances were computed by subtracting the estimated overall mean from the actual overall mean and by subtracting the estimated overall variances from the actual variance for the saline and anorectic drug conditions.

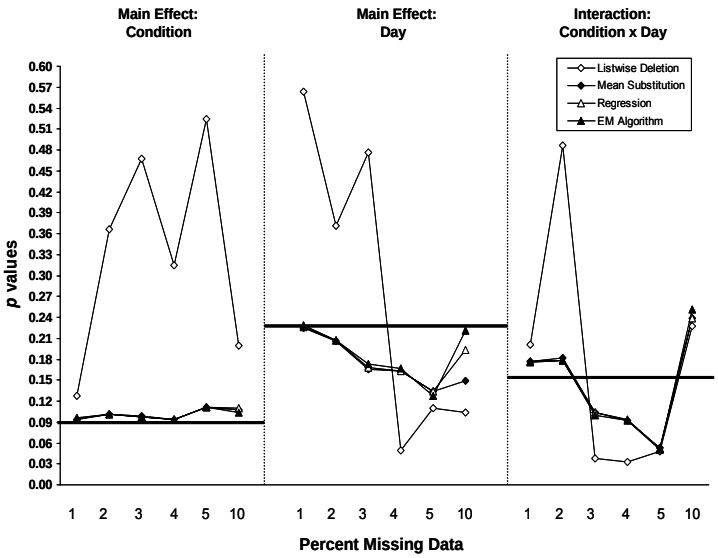

Figure 1. *P*-values plotted as a function of percent of missing data and missing data technique. The black horizontal lines mark the *p*-values for the complete dataset. *P*-values falling below the black line could constitute a Type I error and *p*-values falling above the black line could constitute a Type II error.

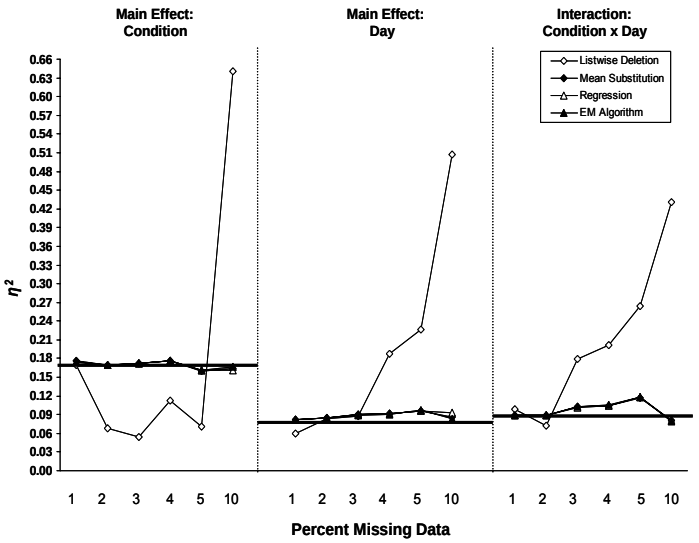

Figure 2. Effect sizes (η^2) plotted as a function of percent of missing data and missing data technique. The black horizontal lines mark the effect size (η^2) for the complete dataset.

Listwise deletion was the least effective method. Listwise deletion tended to either provide more or less support to the null hypothesis. Specifically, listwise deletion resulted in the null hypothesis (no differences between Conditions) receiving more support for the main effect of Condition. However, for the main effect of Day and for the Condition x Day interaction there was a vacillation between providing more or less support for the null hypotheses. With 1% missing data, all methods were effective for the Bayesian approach. As the percentage of missing data increased, all methods gave more support to the null hypothesis for Condition, less support for the null hypothesis for Day, and wavering support for the Condition x Day interaction.

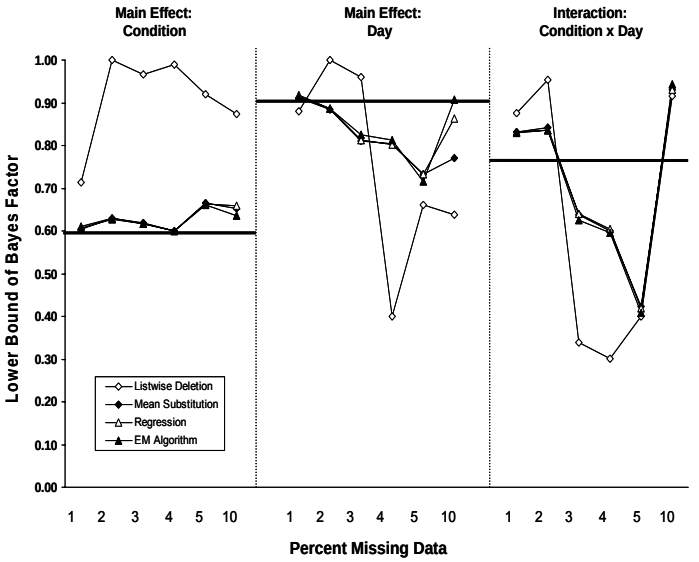

Figure 3. Lower bound of the Bayes factor plotted as a function of percent of missing data and missing data technique. The Black horizontal lines mark the lower bound of the Bayes factor for the complete dataset.

DISCUSSION

We investigated methods for handling missing data on a complete dataset that examined the effects of feeding behaviors between drug-treated and untreated animals using frequentist and Bayesian data analytic techniques. For this sample and as expected, listwise deletion was the least efficacious method for the mixed design ANOVAs (frequentist approach) and Bayes factors (Bayesian approach). With 1 or 2% missing data, mean substitution, regression, and EM were efficient procedures according to *p*-values and effect sizes (frequentist approach) and Bayes factors (Bayesian approach). As the percent of missing data increased to 10%, the frequentist and Bayesian approaches showed that EM had a slight advantage over regression and mean substitution and regression had a small advantage over mean substitution. In this study EM and regression were the preferred methods for handling the missing data with small samples with 10% missing data.

Our analyses showed, as expected, that listwise deletion is not an effective method for computing mixed design ANOVAs when greater than 5% of the data are missing. This finding has serious implications given that researchers in the behavioral neurosciences tend to rely on listwise deletion in conducting complete case analysis, where only samples with complete information on all variables of interest are included in the analysis and cases with missing data are deleted. Although a convenient method, this procedure makes the assumption that samples with missing data are a random subset from the overall sample of data and that the missing data are MCAR. However, if the subset of samples is not a representative sample from the entire dataset then the probability of inaccurate results increases. Additionally, the samples with missing data might be MAR or NIM. For instance, missing pellet consumption in the anorectic drug condition might be a result of obese rats and listwise deletion would ignore these animals. Consequently, the amount of pellet consumption would be an overestimation of the true pellet consumption in the sample.

Regression and EM were the preferred estimation methods compared to mean substitution especially in the case of 10% missing data. It is important to note that the adequacy of regression as a missing data method is dependent upon the predictors selected for the regression equation. Therefore, regression would not be the missing data method of choice with poor study predictors or only one predictor. Although EM also relies on selected predictors, EM is an iterative estimation procedure which reduces the impact of poor predictors by maximizing the expected values across all predictors.

There were several limitations to the present study. First, we used real data rather than conducting a simulation study, which would allow us to test differences across missing data procedures while controlling for effect size, sample size, and variable distributions. We chose to use real data because we wanted to have an externally valid example that would be representative of data commonly encountered in behavioral research. However, one weakness of using real data is that the results could be

specific to idiosyncrasies in the data set and/in sampling or are reflective of theoretical expectations. Second, these results are specific to researchers using small samples and conducting ANOVAs (between subjects, repeated, or mixed design) on the data. Third, we did not compare multiple imputation or maximum likelihood (the two most preferred methods for handling missing data) with the current methods due to our small sample size and the added software requirements to use these methods. Given the aforementioned limitations replications are important as well as the need for generalizations to other statistical analyses (i.e., linear, multivariate, or mixed effects regression) and sample sizes. With the recent advancements in missing data methods researchers are able to move beyond ignoring missing data or mean substitution methods. For any multivariate analysis deleting cases is an inefficient method, especially when sample sizes are small and there are a large percent of missing data.

REFERENCES

Allison, PD (2002) Missing data. Thousand Oaks, CA: Sage Publication.

Bernardo JM, Smith AFM (1994) Bayesian theory. Hoboken, NJ: Wiley.

Choi YJ, Nam CM, Kwak MJ (2004) Multiple imputation technique applied to appropriateness ratings in cataract surgery. Yonsei Med J 45:829-837.

Cohen J (1994) The earth is round (p <.05). Am Psychol 49:997-1003.

Collins LM, Schafer JL, Kam CM (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 6:330-351.

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1-38.

Enders C, Peugh J (2004) Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. Structural Equation Modeling: A Multidisciplinary Journal 11:1-19.

Enders CK (2001) A primer on maximum likelihood algorithms available for use with missing data. Structural Equation Modeling: A Multidisciplinary Journal 8:128-141.

Evenden JL (1999) The pharmacology of impulsive behaviour in rats VII: The effects of serotonergic agonists and antagonists on responding under a discrimination task using unreliable visual stimuli. Psychopharmacology (Berl.) 146:422–431.

Gonzales RA, Weiss F (1998) Suppression of ethanol-reinforced behavior by Naltrexone is associated with attenuation of the ethanol-induced increase in dialysate dopamine levels in the nucleus accumbens. J Neurosci 18:10663-10671.

Greenhouse SW, Geisser S (1959) On methods in the analysis of profile data. Psychometrika 32:95-112.

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

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773-795.

Little RJA (1979) Maximum likelihood inference for multiple regression with missing values: A simulation study. J R Stat Soc Ser B (Methodological) 41:76-87.

Little RJA, Rubin DB (1987) Statistical analysis with missing data. New York, NY: Wiley.

Little RJA, Rubin DB (1989) The analysis of social science data with missing values. Sociol Methods Res 18:292-326.

Raudenbush SW, Bryk AS, (2002) Hierarchical linear models: Application and data analysis methods. Thousand Oaks, CA: Sage Publications.

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

Rubin DB (1976) Inference and missing data. Biometrika 63:581-592.

Rubin, DB (1977) Formalizing subjective notion about the effect of nonrespondents in sample surveys. J Am Stat Assoc 72:538-543.

Schafer JL (1999) Multiple imputation: A primer. Stat Methods Med Res 8:3-15.

Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. Psychol Methods 7:147-177.

Sellke T, Bayarri M, Berger J (2001) Calibration of p values for testing precise null hypotheses. Am Stat 55:62-71.

Sokoloff G, Blumberg MS (2001) Competition and cooperation among huddling infants. Dev Psychobiol 39:65-75.

Sokoloff G, Blumberg MS, Adams MM (2000) A comparative analysis of huddling in infant Norway rats and Syrian golden hamsters: Does endothermy modulate behavior? Behav Neurosci 114:585-593.

Tkacs NC, Li J, Strack AM (1997) Central amygdala Fos expression during hypotensive or febrile, nonhypotensive endotoxemia in conscious rats. J Comp Neurol 379:592-602.

Wothke W (1998) Longitudinal and multi-group modeling with missing data. In: Modeling longitudinal and multiple group data (Little T, Schnabel K, Baumert J, eds). Mahwah, NJ: Lawrence Erlbaum Associates.

Received March, 08, 2007; revised April, 18, 2007; accepted April, 20, 2007.

Original data collection was supported by grants DC04341 and DC06456 from the National Institute of Deafness and Other Communication Disorders to SR. The authors are grateful to Drs. George Karabatsos and Linda Skitka for helpful discussions and suggestions on previous versions of this manuscript.

Address correspondence to: Leah H. Rubin, Center for Cognitive Medicine, Neuropsychiatric Institute, University of Illinois at Chicago, 912 S. Wood St., Suite 235, Chicago, IL 60612. Email: lrubin@psych.uic.edu