Journal of Modern Applied Statistical Methods

Volume 17 | Issue 1 Article 5

2018

Handling Missing Data in Single-Case Studies

Chao-Ying Joanne Peng
Indiana University Bloomington, peng@indiana.edu

Li-Ting Chen
University of Nevada, Reno, litingc@unr.edu

Recommended Citation
Peng, C-Y. J., & Chen, L-T. (2018). Handling Missing Data in Single-Case Studies. Journal of Modern Applied Statistical Methods, 17(1), eP2488. doi: 10.22237/jmasm/1525133280

Handling Missing Data in Single-Case Studies

Cover Page Footnote
We thank Po-Ju Wu for his insight into literature on missing data methods and careful reading of this paper.

Journal of Modern Applied Statistical Methods
May 2018, Vol. 17, No. 1, eP2488
doi: 10.22237/jmasm/1525133280

ISSN 1538 − 9472

Handling Missing Data in Single-Case Studies

Chao-Ying Joanne Peng
Indiana University Bloomington
Bloomington, IN

Li-Ting Chen
University of Nevada, Reno
Reno, NV

Multiple imputation is illustrated for dealing with missing data in a published SCED study. Results were compared to those obtained from available data. Merits and issues of implementation are discussed. Recommendations are offered on primal/advanced readings, statistical software, and future research.

Keywords: Missing data, single case, imputation, intervention effect

Introduction

The occurrence of missing data is prevalent in single-case experimental design (SCED) studies due to the repeated observation and assessment of an outcome behavior in such settings (Franklin, Allison, & Gorman, 1996). Smith (2012) reviewed SCED standards and 409 studies published in refereed journals between 2000 and 2010, and noted “SCEDs undeniably present researchers with a complex array of methodological and research design challenges, such as establishing a representative baseline,…and appropriately addressing the matter of missing observations” (p. 511). Similarly, articles published from 2015 to summer 2016 in five journals (*Behavior Modification*, *Journal of Applied Behavior Analysis*, *Journal of Positive Behavior Interventions*, *Journal of School Psychology*, and *The Journal of Special Education*) with similar aims to publish behavioral analysis studies in clinical and school settings were examined, and 34 (24%) contained missing data. Another 10 (7%) had insufficient information to determine whether missing data existed.

According to the review by Chen, Peng, and Chen (2015) of computing tools suitable for analyzing SCED data, missing data are commonly handled in one of three ways: (a) deleting missing data via, e.g., RegRand (http://www.matt-

doi: 10.22237/jmasm/1525133280 | Accepted: June 15, 2017; Published: June 7, 2018.
Correspondence: Li-Ting Chen, litingc@unr.edu

koehler.com/regrand), (b) omitting missing sessions or intervals, thus yielding results based on available data only (e.g., PROC MEANS in SAS); or (c) replacing missing data with 0 (e.g., Simulation Modeling Analysis). Unfortunately, these approaches waste information already collected, distort the initial SCED design, or misrepresent the results.

Even though treating missing data is usually not the focus of a substantive study, failing to do so properly threatens internal validity, the statistical conclusion validity, and weakens the generalizability of any SCED study (Rubin, 1987; Schafer, 1997; Shadish, Cook, & Campbell, 2002). Serious consequences can result from improper treatments of missing data. First, deleting cases or sessions with missing data listwise leads to the loss of information which wastes information already collected. Furthermore, the reduced sample may not be representative of the population because participants with missing data are not removed randomly. A reduced sample is always associated with decreased statistical power and increased sampling errors. Second, missing data may prevent researchers from fully analyzing the data. Third, removing missing data inevitably alters the study design and creates difficulty in integrating results across participants or studies. Popular *ad hoc* methods, such as mean substitution or personal best guesses, can artificially inflate correlation among scores, introduce trends not supported by data, bias the parameter estimate, and result in inefficient inferences (Little & Rubin, 1987, 2002; Peng, Harwell, Liou, & Ehman, 2006). Even when visual analysis is employed to determine an intervention effect, a linear interpolation may be superimposed over missing data, thus, creating a linear trend that may not really exist. Despite this drawbacks, proper treatment of missing data is vital to the validity of conclusions drawn from visual analysis as well as from statistical inferences of SCED data (Smith, 2012).

In a Monte Carlo simulation study, Smith, Borckardt, and Nash (2012) investigated the statistical power when the Expectation-Maximization (EM) procedure was used to replace missing observations in single case time-series data. They simulated 10 and 56 data points in the baseline phase and the intervention phase, respectively, in order to apply the EM procedure. Although this is a principled approach to treating missing data, it is unrealistic to expect to collect 56 data points from the intervention phase in most SCED studies. Therefore, the results from Smith et al. may not be generalizable.

Because of the threats posed by missing data to interpretations of SCED data by visual analysis or by statistics, missing data should not be ignored and should be treated properly. The objective of this paper is to illustrate another principled method, multiple imputation, as a viable approach for handling missing data in

SCED studies. It can retain the information already collected and allow for valid statistical inferences. Even though multiple imputation has not been routinely employed in SCED to deal with missing data, its rationale can be easily understood given basic statistical knowledge. Thus, this paper aims to (1) illustrate multiple imputation by applying this approach systematically to missing data in a published SCED study (Lambert, Cartledge, Heward, & Lo, 2006), and (2) discuss practical issues surrounding the application of multiple imputation in SCED contexts. Six features of the Lambert et al. (2006) data, including level/level change, trend, variability, immediacy of the effect, overlap, and consistency of data in similar phases, were systematically assessed, according to the recommendations of the What Works Clearinghouse (WWC) Standards Handbook (WWC, 2017), hereafter abbreviated as the WWC Handbook and its standards as the WWC Standards. Results were contrasted with those based on available data in order to determine the effectiveness of an intervention effect. All assessments were conducted using SAS 9.4 or the Single Case Research (SCR) website (http://www.singlecaseresearch.org).

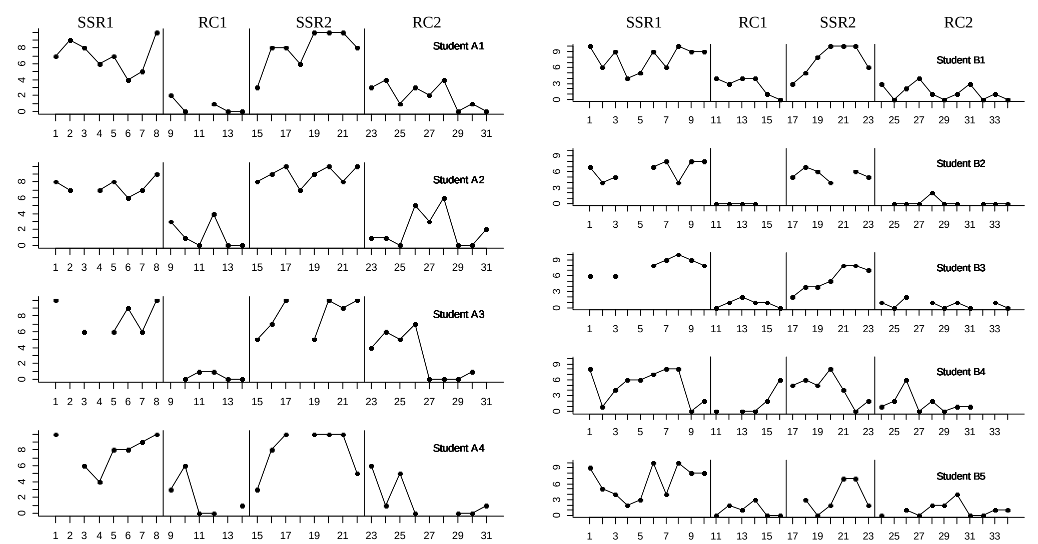

Figure 1. Number of intervals of disruptive behaviors during single-student responding (SSR) and response card (RC) conditions; adapted from Lambert et al. (2006)

The Lambert Data Set

Lambert et al. (2006) implemented a strategy, namely, the response cards or RC, to minimize students' disruptive behaviors during math instruction. The study was conducted in two fourth-grade classrooms with a total of nine target students from a Midwestern urban elementary school. A reversal (or an ABAB) design was employed with two baseline phases (SSR1 and SSR2) and two intervention phases (RC1 and RC2). A disruptive behavior, such as engaging in a conversation during teacher-directed instruction, provoking others, laughing, or touching others, was recorded in 10 intervals of each study session. The dependent variable was the number of intervals in which at least one disruptive behavior was observed, with 10 as the maximum and 0 as the minimum (see Figure 1). Using visual analyses and analysis of means, Lambert et al. concluded that the use of response cards was effective in decreasing disruptive behaviors for these nine students.

The breaks in Figure 1 were due to student absences. All students, except for B1, had missing data with an average missing data rate at 10%. The highest missing rate was 7 (23%) for A3. These breaks were ignored in the analyses of Lambert et al. (2006) and in the reanalysis of these data, published in volume 52, issue 2 of the *Journal of School Psychology*, to demonstrate alternative ways of analyzing SCED data beyond visual analysis. They were acknowledged as missing data in Chen et al. (2015) and in Peng and Chen (2015) in order to maintain the initial structure of this study design, while evaluating computing tools' accuracy and treatment of missing data for analyzing SCED data.

Multiple Imputation

Multiple imputation provides valid statistical inferences under the missing at random (MAR) condition (Little & Rubin, 2002). The MAR condition assumes that the probability of missing is not a function of the missing scores themselves, but may be a function of observed scores (Little & Rubin, 1987). The MAR condition can be made more plausible if variables that explain missingness are included in the statistical inferential process.

Multiple imputation imputes missing data while accounting for the uncertainty associated with the imputed values (Little & Rubin, 2002). It accounts for the uncertainty by generating a set of *m* plausible values for each unobserved data point, resulting in *m* complete data sets, each with an estimate of the missing values. The *m* plausible values represent a random sample of all plausible values given the observed data. The *m* complete data sets are analyzed separately using a standard statistical procedure, resulting in *m* slightly different estimates for each parameter, such as the mean. At the final stage of multiple imputation, *m* estimates are pooled together to yield a single estimate of the parameter and its corresponding *SE*. The pooled *SE* of the parameter estimate incorporates the uncertainty due to the missing data treatment (the between imputation uncertainty) into the uncertainty inherent in any estimation method (the within imputation uncertainty). Consequently, the pooled *SE* is larger than the *SE* derived from a single imputation method (e.g., mean substitution) that does not consider the between imputation uncertainty. Thus, multiple imputation can more accurately estimate the *SE* of a parameter estimate than a single imputation method (Little & Rubin, 2002; Rubin, 1987). Multiple imputation treats missing data in three steps: (1) imputes missing data *m* times to produce *m* complete data sets; (2) analyzes each data set using a standard statistical procedure; and (3) pools, or combines, the *m* results into one using formulae from Rubin (1987).

Step 1: Imputation

The imputation step fills in missing values multiple times using the information already contained in the observed data. The preferred imputation method is the one that matches the missing data pattern. Three missing data patterns have been identified in the literature: univariate, monotone, and arbitrary. A data set has a univariate pattern of missing data if the same participants have missing data on the same variable(s), or in the same sessions as in Lambert et al. (2006). A dataset has a monotone missing data pattern if the variables, or sessions, can be arranged in such a way that, when one variable/session score is missing, the subsequent variables/session scores are missing as well. The monotone missing data pattern occurs frequently in longitudinal studies where, if a participant drops out at one point, their data are missing on subsequent measures. If missing data occur in any variable/session for any participant in a random fashion, the data set is said to have an arbitrary missing data pattern.

Given a univariate or monotone missing data pattern, one can impute missing values using the regression method (Rubin, 1987), or using the predictive mean matching method if the missing variable is continuous (Heitjan & Little, 1991; Schenker & Taylor, 1996). When the missing data pattern is arbitrary, one can use the Markov Chain Monte Carlo (MCMC) method (Schafer, 1997), or the fully conditional specification (FCS, also referred to as chained equations) if the missing variable is categorical or non-normal (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren, 2007; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). MCMC assumes the joint distribution for all variables in the imputation model to be multivariate normal, or bivariate normal if there are two variables in the imputation model. FCS does not hold this normality assumption. Both MCMC and FCS have been implemented in PROC MI (SAS Institute Inc., 2015).

For the Lambert data set, the missing mean of each phase was imputed for all students with a missing score in that phase because there were different numbers of sessions implemented in Classes A and B. The MCMC method was specified based on the arbitrary pattern of the missing means (meanssr1 to meanrc2, highlighted in grey in Table 1) and continuous variables included in the imputation models. Four imputation models were constructed, each composed of two variables: the mean variable with missing data and an auxiliary variable with complete data. The mean variable and the auxiliary variable were from the same phase and were strongly correlated; the absolute correlations ranged from 0.31 to 0.58. To impute missing SSR1 means (or meanssr1 in Table 1), the auxiliary variable, ar_ssr1, was included in the imputation model. The variable ar_ssr1 was academic response during the SSR1 phase. Academic response was defined as "an observable response made by the student to the teacher's question. In this study, an academic response was scored when a student orally responded to the teacher's instructional question after raising his or her hand and being called on (during single-student responding), or when he or she wrote down the answer on the white board following the teacher's question (during response cards) during math lessons" (Lambert et al., 2006, p. 90). Similarly, the auxiliary variable ar_ssr2 was used in imputing the missing SSR2 means (meanssr2). To impute the missing RC1 or RC2 means (meanrc1 or meanrc2), the auxiliary accuracy variables acc_rc1 and acc_rc2 were used, respectively. Acc_rc1 (acc_rc2) measured the percent of times when students gave a correct answer written on the response card during the RC1 (RC2) phase (Lambert et al., 2006, p. 90). The four imputation models had to be simple in order to allow PROC MI to converge, due to the small number of students ($n = 9$).

The auxiliary variables were selected for their strong correlation with the missing means (from the same phase), completeness (no missing data themselves), and for a wide range or variability (to avoid the singularity problem). One set of imputed means are shown in grey in Table 1. SAS computing codes for this step are shown in Part A of Appendix A.

At the end of Step 1 – Imputation, five sets ($m = 5$) of complete data were generated. Given the average missing data rate at about 10%, five imputations were considered sufficient (Graham & Schafer, 1999). One set of complete data from Class A are shown in Tables 2a (for SSR1-RC1 phases) and 2b (for SSR2-RC2 phases), and one set of complete data from Class B are shown in Tables 3a (for SSR1-RC1 phases) and 3b (for SSR2-RC2 phases).

Table 1. Four auxiliary variables (ar_ssr1 to acc_rc2) and one set of imputed scores for four variables with missing data (meanssr1 to meanrc2)

Student	ar_ssr1	acc_rc1	ar_ssr2	acc_rc2	meanssr1	meanrc1	meanssr2	meanrc2
A1	0.11	97.1	0.13	95.7	7.0000	0.9819	7.8750	2.0000
A2	0.10	95.2	0.05	96.7	6.1405	1.3333	8.8750	2.0000
A3	0.08	68.5	0.03	96.7	7.3748	5.2679	8.5762	2.0293
A4	0.00	93.8	0.07	96.4	9.0090	1.5716	9.7142	2.0277
B1	0.10	91.8	0.03	81.1	7.7000	2.6666	7.4286	1.3636
B2	0.18	85.7	0.15	97.8	4.8478	2.9150	2.8326	2.0739
B3	0.13	100.0	0.05	72.6	5.8749	0.8333	5.4286	1.0088
B4	0.17	90.0	0.10	94.6	5.0000	2.1988	4.2857	1.9764
B5	0.07	97.1	0.08	90.8	6.3000	1.0000	1.8793	1.7918

Note: Only one set of imputed means (meanssr1 to meanrc2) are shown in grey highlights here; variables meanssr1 to meanrc2 are rounded off to four decimal places to preserve precision

Step 2: Analysis

The second step of multiple imputation analyzes the five (or *m*) complete data sets separately using a statistical procedure that was suited for testing differences between phase means, trends, variabilities, and nonoverlap between adjacent phases in SCED data. At the end of the second step, five sets of results were obtained from separate analyses of the five data sets. SAS computing codes for performing *t*-tests and for computing means for each of five imputed data sets are shown in Parts B and C of Appendix A.

Step 3: Pooling

The third step of multiple imputation combines the five (or *m*) results into one. This step is implemented into PROC MIANALYZE. PROC MIANALYZE is useful for pooling results that are obtained from a model-based analysis, such as a regression or logistic analysis. Otherwise, results are pooled using Rubin's formulae in PROC MI (1987) or by hand calculations. Rubin's formulae combine five results and *SE*s into a single result and its *SE*. Suppose $\hat{Q}_i$ denotes the estimate of a parameter Q, (e.g., a mean) from the i^{th} imputed data set. Its corresponding estimated variance is denoted as $\hat{U}_i$. Then the pooled point estimate of Q is given by

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{1}$$

Table 2a. Class A's number of intervals of disruptive behaviors and their ranks from the SSR1 to the RC1 phases (Lambert et al., 2006)

Class A	SSR1								RC1					
Student	**1[a]**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A1	7	9	8	6	7	4	5	10	2	0	0.9819	1	0	0
A2	8	7	6.1405	7	8	6	7	9	3	1	0	4	0	0
A3	10	7.3748	6	7.3748	6	9	6	10	5.2679	0	1	1	0	0
A4	10	9.0090	6	4	8	8	9	10	3	6	0	0	1.5716	1

	SSR1-Ranks								RC1-Ranks					
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A1	10.5	13.0	12.0	9.0	10.5	7.0	8.0	14.0	6.0	2.0	4.0	5.0	2.0	2.0
A2	12.5	10.0	8.0	10.0	12.5	7.0	10.0	14.0	5.0	4.0	2.0	6.0	2.0	2.0
A3	13.5	10.5	8.0	10.5	8.0	12.0	8.0	13.5	6.0	2.0	4.5	4.5	2.0	2.0
A4	13.5	12.0	7.5	6.0	9.5	9.5	11.0	13.5	5.0	7.5	1.5	1.5	4.0	3.0
Total rank	50.0	45.5	35.5	35.5	40.5	35.5	37.0	55.0	22.0	15.5	12.0	17.0	10.0	9.0
Expected rank[b] (Y_j)	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Note: Values in grey highlights in the upper panel were missing in Lambert et al. and imputed in this study using multiple imputation; their corresponding ranks are in grey in the lower panel. Only one set of imputed values are presented here.

[a] Session numbers

[b] Expected ranks are derived from H_1 of the Page test

Table 2b. Class A's number of intervals of disruptive behaviors and their ranks from the SSR2 to the RC2 phases of class A (Lambert et al., 2006)

Class A	**SSR2**								**RC2**								
Student	**15[a]**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
A1	3	8	8	6	10	10	10	8	3	4	1	3	2	4	0	1	0
A2	8	9	10	7	9	10	8	10	1	1	0	5	3	6	0	0	2
A3	5	7	10	8.5762	5	10	9	10	4	6	5	7	0	0	0	1	2.0293
A4	3	8	10	9.7142	10	10	10	5	6	1	5	0	2.0277	2.0277	0	0	1

	SSR2-Ranks								**RC2-Ranks**								
Student	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
A1	7.0	13.0	13.0	11.0	16.0	16.0	16.0	13.0	7.0	9.5	3.5	7.0	5.0	9.5	1.5	3.5	1.5
A2	11.5	13.5	16.0	10.0	13.5	16.0	11.5	16.0	4.5	4.5	2.0	8.0	7.0	9.0	2.0	2.0	6.0
A3	8.0	11.5	16.0	13.0	8.0	16.0	14.0	16.0	6.0	10.0	8.0	11.5	2.0	2.0	2.0	4.0	5.0
A4	8.0	12.0	15.5	13.0	15.5	15.5	15.5	9.5	11.0	4.5	9.5	2.0	6.5	6.5	2.0	2.0	4.5
Total rank	34.5	50.0	60.5	47.0	53.0	63.5	57.0	54.5	28.5	28.5	23.0	28.5	20.5	27.0	7.5	11.5	17.0
Expected rank[b] (Y_j)	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Note: Values in grey highlights in the upper panel were missing in Lambert et al. and imputed in this study using multiple imputation; their corresponding ranks are in grey in the lower panel. Only one set of imputed values are presented here.

[a] Session numbers

[b] Expected ranks are derived from H_1 of the Page test

Table 3a. Class B's number of intervals of disruptive behaviors and their ranks from the SSR1 to the RC1 phases (Lambert et al., 2006)

	SSR1										RC1					
Student	**1[a]**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
B1	10	6	9	4	5	9	6	10	9	9	4	3	4	4	1	0
B2	7	4	5	4.8478	4.8478	7	8	4	8	8	0	0	0	0	2.9150	2.9150
B3	6	5.8749	6	5.8749	5.8749	8	9	10	9	8	0	1	2	1	1	0
B4	8	1	4	6	6	7	8	8	0	2	0	2.1988	0	0	2	6
B5	9	5	4	2	3	10	4	10	8	8	0	2	1	3	0	0

	SSR1-Ranks										RC1-Ranks					
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
B1	15.5	9.5	12.5	5.5	8.0	12.5	9.5	15.5	12.5	12.5	5.5	3.0	5.5	5.5	2.0	1.0
B2	12.5	7.5	11.0	9.5	9.5	12.5	15.0	7.5	15.0	15.0	2.5	2.5	2.5	2.5	5.5	5.5
B3	10.5	8.0	10.5	8.0	8.0	12.5	14.5	16.0	14.5	12.5	1.5	4.0	6.0	4.0	4.0	1.5
B4	15.0	5.0	9.0	11.0	11.0	13.0	15.0	15.0	2.5	6.5	2.5	8.0	2.5	2.5	6.5	11.0
B5	14.0	11.0	9.5	5.5	7.5	15.5	9.5	15.5	12.5	12.5	2.0	5.5	4.0	7.5	2.0	2.0
Total rank	67.5	41.0	52.5	39.5	44.0	66.0	63.5	69.5	57.0	59.0	14.0	23.0	20.5	22.0	20.0	21.0
Expected rank[b] (Y_j)	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Note: Values in grey highlights in the upper panel were missing in Lambert et al. and imputed in this study using multiple imputation; their corresponding ranks are in grey in the lower panel. Only one set of imputed values are presented here.

[a] Session numbers

[b] Expected ranks are derived from H_1 of the Page test

Table 3b. Class A's number of intervals of disruptive behaviors and their ranks from the SSR2 to the RC2 phases of class A (Lambert et al., 2006)

	SSR2							RC2										
Student	**17[a]**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**
B1	3	5	8	10	10	10	6	3	0	2	4	1	0	1	3	0	1	0
B2	5	7	6	4	2.8326	6	5	2.0739	0	0	0	2	0	0	2.0739	0	0	0
B3	2	4	4	5	8	8	7	1	0	2	1.0088	1	0	1	0	1.0088	1	0
B4	5	6	5	8	4	0	2	1	2	6	0	2	0	1	1	1.9764	1.9764	1.9764
B5	1.8793	3	0	2	7	7	2	0	1.7918	1	0	2	2	4	0	0	1	1

	SSR2-Ranks							RC2-Ranks										
Student	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**
B1	10.0	13.0	15.0	17.0	17.0	17.0	14.0	10.0	2.5	8.0	12.0	6.0	2.5	6.0	10.0	2.5	6.0	2.5
B2	14.5	18.0	16.5	13.0	12.0	16.5	14.5	10.5	4.5	4.5	4.5	9.0	4.5	4.5	10.5	4.5	4.5	4.5
B3	11.5	13.5	13.5	15.0	17.5	17.5	16.0	6.5	2.5	11.5	9.5	6.5	2.5	6.5	2.5	9.5	6.5	2.5
B4	14.5	16.5	14.5	18.0	13.0	2.0	11.0	5.0	11.0	16.5	2.0	11.0	2.0	5.0	5.0	8.0	8.0	8.0
B5	10.0	15.0	3.0	12.5	17.5	17.5	12.5	3.0	9.0	7.0	3.0	12.5	12.5	16.0	3.0	3.0	7.0	7.0
Total rank	60.5	76.0	62.5	75.5	77.0	70.5	68.0	35.0	29.5	47.5	31.0	45.0	24.0	38.0	31.0	27.5	32.0	24.5
Exp rank[b] (Y_j)	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Note: Values in grey highlights in the upper panel were missing in Lambert et al. and imputed in this study using multiple imputation; their corresponding ranks are in grey in the lower panel. Only one set of imputed values are presented here.

[a] Session numbers

[b] Expected ranks are derived from H_1 of the Page test

The variance of $\bar{Q}$, denoted as T in equation (4), is the weighted sum of two variances: the within-imputation variance ($\bar{U}$) and the between-imputation variance (B). Specifically, these three variances are computed as follows:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \tag{2}$$

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \bar{Q}\right)^2 \tag{3}$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B = \text{the variance of } \bar{Q} \tag{4}$$

In equation (4), the ($1 / m$) factor is an adjustment for a lack of randomness associated with a finite number of imputations. Theoretically, estimates derived from multiple imputation with a small m yield larger sampling variances than maximum-likelihood estimates, such as those derived from full information maximum likelihood, because the latter are not impacted by a lack of randomness caused by multiple imputation.

Assessment Results According to the WWC Standards

In assessing the intervention effect, the WWC Standards (WWC, 2017) were used while treating missing data in the Lambert Data Set. These standards were formulated to help researchers and practitioners determine whether (a) the observed pattern of data in the intervention phase is due to the intervention effects, and (b) the observed pattern of data in the intervention phase is different from the pattern of data, predicated from data in the baseline phase. Six data features, both within and between phases, are recommended for analysis by the WWC Standards in order to determine the effectiveness of an intervention effect. Analyses of data collected from the SSR1 to RC1 and the SSR2 to RC2 phases are presented.

Assessment of Level/Level Change

In the WWC Handbook, level was defined as the mean score for data within a phase (WWC, 2017). A level change between phases indicates a change in the outcome measure due to the intervention. To assess the level and level change from SSR to RC phases, the paired-samples t-test was applied to means obtained from one

Table 4. Means, *SD*s, paired-samples *t*-tests of differences between SSR-RC phases in the Lambert data set

	Available Data		Multiple Imputation	
Statistics	**SSR1-RC1**	**SSR2-RC2**	**SSR1-RC1**	**SSR2-RC2**
Mean[a]	5.90	5.02	5.48	4.92
SD[b]	1.22	1.55	1.00	1.51
n[c]	9	9	9	9
t-test[d]	14.54	9.75	16.44	9.76
p-value[e]	<0.0001	<0.0001	<0.0001	<0.0001

Note: [a]Means are computed as an average of individuals' difference score over sessions between SSR and RC phases

[b] SDs are computed as the square root of the variance of individuals' difference scores

[c] n = number of students

[d] paired t-test of SSR−RC differences, $df = 9 - 1 = 8$

[e] one-tailed probability, rounded up to four decimal places for precision and comparisons

baseline and one intervention phases, namely, the means of SSR1-RC1, and the means of SSR2-RC2 using available data and multiple imputation. At Step 3 of multiple imputation, *t*-statistics and *p*-values were pooled across five imputed data sets. Results based on multiple imputation and the Available Data are presented in Table 4.

According to Table 4, the four paired-samples *t*-tests for the difference between one SSR phrase and one RC phrase ranged from 16.44 to 9.75 with $df = 8$ (i.e. 9 – 1). All four paired-samples *t*-tests were statistically significant at $p < 0.0001$ (one-tailed), suggesting a level change, specifically a decline from a SSR phase to an adjacent RC phase. And the decline implied the effectiveness of the intervention. Although the statistical significant results reached the same conclusion based on either the Available Data approach or the Multiple Imputation approach, the means and *SD*s convey a different message. For the two SSR versus RC comparisons, the mean differences obtained from the Multiple Imputation approach (5.48 and 4.92) were not comparable to those under the Available Data approach (5.90 and 5.02), whereas the *SD*s under the Multiple Imputation approach (1.00 and 1.51) were comparable to those under the Available Data approach (1.22 and 1.55).

Assessment of Trend

"Trend refers to the slope of the best-fitting straight line for the data within a phase," according to the WWC Handbook (WWC, 2017, p. A-7). Because a best-fitting straight line is a limiting definition for trends, we elected to assess monotonic trends in the Lambert data sets using the Page test of trends (Busk & Marascuilo, 1992;

Marascuilo & Busk, 1988; Peng & Chen, 2015). A monotonic trend can be either increasing or decreasing. It is more general than a linear trend because a monotonic trend incorporates different slopes throughout a data pattern to reflect an upward (or increasing), or a downward (or decreasing), trend in data. Marascuilo and McSweeney (1977) and Page (1963) recommended the Page test for testing monotonic changes over time. The type of measurement required by the Page test is ranks of data or ranked data. Marascuilo and Busk (1988) and Busk and Marascuilo (1992) applied the Page test to assess trends in the simple AB design, the multiple-baseline AB designs and replicated ABAB designs across students.

The Page test was conducted for all students as well as for their class trends from two adjacent SSR-RC phases. For each Page test, the null hypothesis (H_0) states that there is no trend in data from the SSR phase to the RC phase. The alternative hypothesis (H_1) states that there is at least one monotonic decreasing trend in data, meaning the RC intervention worked. H_0 and H_1 are expressed in ranks of each student's scores. Furthermore, the rejection of H_0 requires at least one inequality, specifically a decline (or improvement). The Page test cannot be conducted when missing data are present, because ranks cannot be assigned to missing scores; consequently, there was no trend assessment based on available data.

Table 5. Page test of trends from SSR1 to RC1 phases for Classes A and B (Lambert et al., 2006)

	Student	Page *L*	χ^2_L	ES = $z = \sqrt{\chi^2_L}$	*z*-lower = *z* – 1.645[a]	One-tailed *p*
Class A	A1	972.8	8.62	2.94	1.29	0.0033
	A2	958.4	7.34	2.71	1.06	0.0068
	A3	953.8	7.00	2.64	0.99	0.0103
	A4	944.0	5.80	2.40	0.77	0.0165
	Aggregate	3829.0	28.97	5.38	3.74	< 0.0001
Class B[b]	B1	1384.0	6.75	2.60	0.95	0.0094
	B2	1326.1	3.81	1.94	0.29	0.0601
	B3	1335.5	4.18	2.04	0.40	0.0412
	B4	1306.0	2.92	1.71	0.06	0.0875
	B5	1360.0	5.40	2.32	0.68	0.0201
	Aggregate	6711.6	22.54	4.75	3.10	< 0.0001

Note: *p*-values are rounded off to four decimal places for comparison purposes; students B1 and B5 had no missing data in SSR1-RC1 phases

[a] This lower limit is for a 95% one-sided CI

[b] Students B3's and B5's Page *L* statistics may not be statistically significant at $p < 0.05$, if between-imputation variability was factored into the SE of Page *L*

Table 6. Page test of trends from SSR2 to RC2 phases for Classes A and B (Lambert et al., 2006)

	Student	Page L	χ^2_L	ES = $z = \sqrt{\chi^2_L}$	z-lower = $z - 1.645$[a]	One-tailed p
Class A	A1	1663.5	7.89	2.81	0.94	0.0050
	A2	1671.0	8.31	2.88	0.73	0.0039
	A3	1639.8	6.64	2.58	1.29	0.0101
	A4	1664.6	8.05	2.82	1.06	0.0079
	Aggregate	6638.9	30.76	5.54	3.90	< 0.0001
Class B[b]	B1	1975.0	8.90	2.98	1.34	0.0029
	B2	1997.8	10.09	3.18	1.53	0.0015
	B3	1967.4	8.52	2.92	1.27	0.0035
	B4	1895.0	5.30	2.30	0.66	0.0213
	B5	1825.3	2.95	1.71	0.06	0.0926
	Aggregate	9660.5	34.27	5.85	4.21	< 0.0001

Note: p-values are rounded off to four decimal places for comparison purposes; A1, A2, and B1 had no missing data in the SSR2-RC2 phases

[a] This lower limit is for a 95% one-sided CI

[b] Student B4's Page L statistic may not be statistically significant at $p < 0.05$, if between-imputation variability was factored into the *SE* of Page L

To apply the Page test, the raw data in the upper panel of Tables 2a, 2b, 3a, and 3b were converted to ranks for each student, shown in the lower panels, based on one set of imputed values. Ranks assigned to imputed values are shown in grey highlights. If scores/imputed values were tied, ties were broken by averaging the two corresponding ranks, such as assigning the rank of 10.5 to the two 7s for Student A1 in Sessions 1 and 5 during the SSR1 phase (Table 2a). A total rank across four students in Class A, or five students in Class B, was subsequently weighted by their expected ranks (Y_j), suggested by the H_1, that is, there was a decreasing trend from SSR to RC adjacent phases. The product of the total rank weighted by its expected rank was summed over all 14 sessions into the Page statistic (L) for Class A and over all 16 sessions for Class B for the SSR1 to RC1 phases shown in Table 5. The Page L statistic was computed for each imputed data set and subsequently pooled across five data sets according to equations (1) to (4). For the SSR2 to RC2 adjacent phases, similar calculations were applied and pooled results are presented in Table 6. The approximate significance p-value of the L statistic was obtained from a chi-square distribution with $df = 1$ (Page, 1963). According to the approximate chi-squares tests, H_0 of no trend was rejected for Classes A and B (Tables 5 and 6) at $p < 0.05$. At the individual level, all four students in Class A exhibited a statistically significant downward trend in the dependent variable, from SSR1 to RC1 phases, and again from SSR2 to RC2 phases. The five students in Class B did not uniformly demonstrate a downward trend due to the RC intervention. Students B2 and B4's Page tests of SSR1 to RC1 data were not statistically significant at $p < 0.05$; Student B5's Page test of SSR2 to RC2 data was not statistically significant at $p < 0.05$ either. The large-sample approximation to the sampling distribution of Page's L statistic yields acceptable Type-I error rates for a directional Page test, as long as the number of sessions > 11 for $\alpha = 0.05$, or the number of sessions > 18 for $\alpha = 0.01$, according to Bradley (1978), Fahoome (2002), and Page (1963).

The L statistic is conceptually and algebraically equivalent to the average Spearman rank correlation coefficient (ρ) between Students' ranked scores and the expected ranks according to a monotonic decreasing (or increasing) trend (Page, 1963; van de Wiel & Di Bucchianico, 2001). The L statistic can therefore be standardized into an ES, or normalized z (Peng & Chen, 2015). These ESs are presented in Table 5 for the SSR1-RC1 phases and Table 6 for the SSR2-RC2 phases. The normalized z is scale-free and ranges from negative to positive values without bounds, like Cohen's d. They differ, however, in their assumptions. Cohen's d assumes normality and equal variances for underlying populations (Cohen, 1988), whereas the normalized z does not, because the latter is based on ranks of data.

Because the normalized z follows a standard normal distribution (e.g., Fahoome, 2002; Lyerly, 1952), a directional 95% CI for the normalized z can be constructed (Peng & Chen, 2015). Tables 5 and 6 show that the lower limits for Classes A and B were positive; the earlier rejection of the H_0 of no or an increasing trend was supported at $p < 0.05$, in favor of a monotonic decreasing trend.

Summary of Page L *tests, ESs and CIs for Page* L

According to Tables 5 and 6, both Classes A and B demonstrated a monotonic decreasing trend from SSR1 to RC1 and from SSR2 to RC2 phases based on the Page tests, the corresponding ESs (or zs), and CIs. At the individual level and $p < 0.05$, Students B2 and B4 did not demonstrate a decreasing trend from SSR1 to RC1, and Student B5 did not demonstrate a decreasing trend from SSR2 to RC2 phases.

Table 7. Means, *SD*s, *t*-tests of differences in similar phases

	Available Data				Multiple Imputation			
Statistics	**SSR1**	**SSR2**	**RC1**	**RC2**	**SSR1**	**SSR2**	**RC1**	**RC2**
Mean[a]	7.05	6.54	1.16	1.52	6.86	6.45	1.42	1.58
SD[b]	2.18	2.36	1.59	1.86	2.24	2.48	1.84	1.75
n[c]	9	9	9	9	9	9	9	9
$\lvert t \rvert$[d]	0.48		0.45		0.36		0.18	
SE	1.07		0.82		1.11		0.85	
p-value (two-tailed)	0.6396		0.6614		0.7208		0.8582	

Note: Missing scores are left as missing under the Available Data condition, replaced by multiply imputed scores under multiple imputation; *p*-values are rounded off to four decimal places for comparison purposes

[a] Means are computed as an average of individuals' mean scores over sessions within each phase

[b] *SD*s are computed as the square root of the averaged variance of individuals' variances of scores within each phase

[c] *n* = number of students

[d] two-tailed *t*-test of equality of two means, *df* = 9 + 9 – 2 = 16

Assessment of Variability

According to the WWC Handbook, "Variability refers to the range or standard deviation of data about the best-fitting straight line" (WWC, 2017, p. A-7). Even though a straight regression line was not fit to the Lambert data, the SD of scores was assessed within and between phases (Table 7). Because Multiple Imputation accounts for uncertainty due to imputation and sampling errors, it introduced greater variability into data than the Available Data approach. Consequently, its corresponding *SD*s were larger than those obtained under the Available Data approach, except for the RC phase, although the differences were no more than 16% of the smaller SD. Under both approaches, the *SD*s were larger for the baseline phases (SSR1 or SSR2) than their corresponding *SD*s of the intervention phases (RC1 or RC2) as the RC treatment had helped to reduce the disruptive behaviors in general.

Assessment of Immediacy of the Effect

According to the WWC Handbook,

> Immediacy of the effect refers to the change in level between the last three data points in one phase and the first three data points of the next. The more rapid (or immediate) the effect, the more convincing the inference that change in the outcome measure was due to manipulation of the independent variable. (WWC, 2017, p. A-7)

Applying this definition to Figure 1 using the visual analysis of the available data, it was determined data patterns in RC1 or RC2 phase exhibited an immediate decreasing effect on disruptive behaviors, compared to data patterns in the corresponding SSR1 or SSR2 phase for all students, except for Student B4 from SSR2 to RC2 phases. For Students A1, A3, B2, B4, and B5, who had at least one missing score among the last three data points of a SSR phase, or the first three data points of a RC phase, only Student A3's imputed scores for RC1, ranging from 5 to 9 across five imputations, did not support the immediacy effect due to the RC intervention. Thus, it was concluded there was an immediacy effect due to the intervention for all students in Classes A and B, except for Students A3 (for the SSR1-RC phases) and B4 (for the SSR2 to RC2 phases).

Assessment of Overlap

According to the WWC Handbook,

> Overlap refers to the proportion of data from one phase that overlaps with data from the previous phase. The smaller the proportion of overlapping data points (or conversely, the larger the separation), the more compelling the demonstration of an effect. (WWC, 2017, p. A-7)

To assess the degree of overlap between SSR and RC phases, Tau-U was computed. It was selected, among a myriad of nonoverlap indices, due to its straightforward interpretation and statistical properties (Parker, Vannest, Davis, & Sauber, 2011). The greater the Tau-U value, the less overlap between SSR and RC phases and, hence, the stronger evidence for an effective intervention effect. The computation of Tau-Us was facilitated by the SCR website. Tau-Us for Classes A and B were computed according to the recommendation of Parker and Vannest (2012), that is, both were weighted averages of individual students' Tau-Us. Results are presented in Table 8 after pooling across five imputed data sets using equations (1)-(4).

According to Table 8, Tau-U for SSR1-RC1 phases in Class A was 0.98, $p < 0.0001$, based on Available Data. This Tau-U is interpreted as 98% of pairs of data formed from SSR1 phase and RC1 phase showed improvement, i.e. declining in the number of intervals in which a disruptive behavior was observed. Tau-U decreased to 0.96 based on Multiple Imputation, still statistically significant at $p < 0.0001$. Tau-U for SSR1-RC1 phases in Class B was 0.90, $p < 0.0001$ based on Available Data. Tau-U decreased to 0.89 based on Multiple Imputation, still statistically significant at $p < 0.0001$. Tau-U for Class A's SSR2-RC2 phases was

0.90, $p < 0.0001$, based on Available Data and Multiple Imputation approach. Tau-U for Class B's SSR2-RC2 phases was 0.82 based on Available Data and 0.83 based on Multiple Imputation. Both Tau-Us were statistically significant at $p < 0.0001$.

Table 8. Tau-U for SSR1-RC1 and SSR2-RC2 based on Available Data and the Multiple Imputation Approach

		Available Data				Multiple Imputation			
	Student	**Tau-U**	**Var**	***z***	***p***	**Tau-U**	**Var[a]**	***z***	***p***
SSR1-RC1	A1	1.0000	0.1167	2.9277	0.0017	1.0000	0.1041	3.0984	0.0010
	A2	1.0000	0.1111	3.0000	0.0014	1.0000	0.1041	3.0984	0.0010
	A3	1.0000	0.1333	2.7386	0.0031	0.9167	0.1197	2.6493	0.0040
	A4	0.9143	0.1238	2.5984	0.0047	0.9292	0.1045	2.8741	0.0020
	Class A	0.9791	0.0302	5.6372	<0.0001	0.9630	0.0269	5.8685	<0.0001
	B1	0.9500	0.0944	3.0913	0.0010	(No missing, same results as Available Data)			
	B2	1.0000	0.1354	2.7175	0.0033	0.9467	0.1008	2.9818	0.0014
	B3	1.0000	0.1111	3.0000	0.0014	1.0000	0.0944	3.2540	0.0006
	B4	0.6400	0.1067	1.9596	0.0250	0.6000	0.0944	1.9524	0.0254
	B5	0.9333	0.0944	3.0370	0.0012	(No missing, same results as Available Data)			
	Class B	0.9018	0.0213	6.1788	<0.0001	0.8852	0.0191	6.4014	<0.0001
SSR2-RC2	A1	0.9167	0.0833	3.1754	0.0007	(No missing, same results as Available Data)			
	A2	1.0000	0.0833	3.4641	0.0003	(No missing, same results as Available Data)			
	A3	0.8036	0.0952	2.6039	0.0046	0.8083	0.0845	2.7807	0.0027
	A4	0.8571	0.1020	2.6833	0.0036	0.8417	0.0974	2.6973	0.0035
	Class A	0.8993	0.0226	5.9866	<0.0001	0.9033	0.0256	5.6447	<0.0001
	B1	0.9481	0.0823	3.3057	0.0005	(No missing, same results as Available Data)			
	B2	1.0000	0.0988	3.1820	0.0007	1.0000	0.0823	3.4868	0.0002
	B3	0.9841	0.0899	3.2814	0.0005	1.0000	0.0823	3.4868	0.0002
	B4	0.5536	0.0952	1.7938	0.0364	0.5974	0.0823	2.0830	0.0186
	B5	0.5667	0.0945	1.8439	0.0326	0.6156	0.0837	2.1281	0.0167
	Class B	0.8150	0.1355	6.0153	<0.0001	0.8329	0.0165	6.4816	<0.0001

Note: All numbers are rounded off to four decimal places for comparison purposes; when combining the results for Class A and Class B, Tau-U is weighted by the inverse of the variance. The standard errors of Classes A and B are computed from $\sqrt{\left[1/\left(1/\text{Var}A1+1/\text{Var}A2+1/\text{Var}A3+1/\text{Var}A4\right)\right]}$ and $\sqrt{\left[1/\left(1/\text{Var}B1+1/\text{Var}B2+1/\text{Var}B3+1/\text{Var}B4\right)\right]}$, respectively. This method to combine the results from multiple students is suggested by Parker and Vannest (2012)

[a] Variance of Tau-U is an inverse function of the total number of sessions or intervals in Class A (SSR1 = 8, RC1 = 6, SSR2 = 8, and RC2 = 9) or in Class B (SSR1 = 10, RC1 = 6, SSR2 = 7, and RC2 = 11)

All students demonstrated statistically significant improvement from SSR1 phase to RC1 phase and from SSR2 phase to RC2 phase at $p < 0.05$, based on the two approaches. Variances of Tau-Us were expected to decrease from the Available Data condition to Multiple Imputation condition, indeed they decreased for all students in Table 8, because Tau-U's variance is an inverse function of the data points; the more data points, the smaller the variance.

Assessment of consistency of data in similar phases

According to the WWC Handbook,

> Consistency of data in similar phases involves looking at data from all phases within the same condition… and examining the extent to which there is consistency in the data patterns from phases with the same conditions. The greater the consistency, the more likely the data represent a causal relation. (WWC, 2017, p. A-7)

To determine the consistency of data, we employed the visual analysis of the data patterns between SSR1 and SSR2, and between RC1 and RC2 phases. Furthermore, four independent-samples *t*-tests were applied to similar phases (SSR1 vs. SSR2, and RC1 vs. RC2). According to Table 7, the *t*-test was not statistically significant for either the baseline (SSR1 vs. SSR2) or the intervention (RC1 vs. RC2) phases at $p < 0.05$ under the Available Data and Multiple Imputation conditions. These results suggested that the mean scores obtained from similar phases were not statistically different from each other. We concluded that there was insufficient evidence to imply inconsistency of data patterns within SSR or RC phases, regardless of how missing data were ignored by the Available Data approach, or treated by Multiple Imputation.

Summary Based on Six Assessments

The analyses summarized in Tables 2-8 and interpreted above examined all data features recommended by the WWC Handbook (WWC, 2017) for determining an intervention effect. These assessments showed that the intervention worked between each SSR phase and its adjacent RC phase for Class A and Class B as groups. At the student level, Student B1 was the only student with complete data, whose disruptive behaviors decreased significantly from baseline to intervention phases. The disruptive behavior for Students A1, A2, A3, and A4 decreased significantly during SSR1-RC1 and SSR2-RC2. Students B2's and B4's disruptive behaviors did not decrease significantly at $p < 0.05$ during the SSR1-RC1 phases, but did so during the SSR2-RC2 phases. Student B5's disruptive behaviors decreased significantly at $p < 0.05$ during the SSR1-RC1 phases, but not during SSR2-RC2 phases. In terms of Tau-U as a nonoverlap index, all students exhibited significant nonoverlap between SSR-RC phases based on available data or imputed data. The detailed analyses of individual's trends (Tables 5 and 6) and nonoverlap (Table 8) complemented results reported in Lambert et al. (2006). According to Lambert et al. (p. 93), two students (A2 and B2) showed no overlapping data and three students (A1, B1, and B3) showed only one overlapping data point between the SSR and RC phases. No other data features were analyzed for individual students by Lambert et al. This assessment enriched the analysis of information collected from the nine students and provided evidence that the linear interpolation approach for handling missing data, such as the missing score for Student A3 during the RC1 phase, did not agree with imputed scores, ranging from 5 to 9, according to multiple imputation.

Conclusion

Missing data occur in various patterns and to varying degrees (Cohen & Cohen, 1983, pp. 275-299). About 24% of SCED studies published in the five journals from 2015 to summer of 2016 clearly had missing data. Because serious consequences (e.g., threat to internal validity and statistical conclusion validity, limited generalizability, loss of information, bias) can result from improper treatment of missing data in visual analysis as well as in statistical inferences, this paper aims to propose and illustrate multiple imputation as a principled method for dealing with missing data in SCED studies. Multiple imputation was applied to the Lambert data set (Lambert et al., 2006) in three steps: imputation, analysis, and pooling. Five imputed data sets were analyzed separately, then pooled according to the rules of Robin (1987) for each assessment. Results derived from multiple imputation suggested that the RC intervention worked effectively for both classes. However, individual students did not uniformly benefit from this intervention, as previously noted in the section of Summary Based on Six Assessments. The analyses of individual data enrich and complement the findings reported in Lambert et al. (2006).

The missing data phenomenon shown in Students A1-A4, and B2-B5 of the Lambert data set is referred to as item-level missing (Dong & Peng, 2013). The impact of item-level missing on the validity of research findings depends on the mechanisms that led to missing data, the pattern of missing data, and the proportion of data missing (Dong & Peng, 2013; Tabachnick & Fidell, 2001, p. 58). All are relevant concepts to the understanding of multiple imputation and its implementation. Multiple imputation assumes that missing data mechanism is MAR (or missing at random). Given only the observed data, it is impossible to test whether the MAR condition holds (Carpenter & Goldstein 2004; Horton & Kleinman, 2007; White, Royston, & Wood, 2011). The plausibility of MAR can be examined by a simple *t*-test of mean differences between the group with complete data and that with missing data (Diggle, Heagerty, Liang, & Zeger, 1995; International Business Machines Corporation, 2011; Tabachnick & Fidell, 2013). Variables may be included in the statistical inferential process that could explain missingness to make the MAR condition more plausible.

Multiple imputation is applicable to any pattern of missing data, whether it is univariate, monotone, or arbitrary (Little & Rubin, 1987; Rubin, 1987). Regarding an acceptable proportion of missing data for valid statistical inferences, there is no established cutoff in the literature, even though such a proportion directly impacts the quality of statistical inferences. Schafer (1999) asserted that a missing rate of 5% or less is inconsequential. Bennett (2001) maintained that statistical analysis is likely to be biased when more than 10% of data are missing. Dong and Peng (2013) investigated the performance of multiple imputation, against the complete data approach, under 20%, 40%, and 60% of missing data conditions in a between-subject data set. In terms of bias and standard errors of parameter estimates, the 20% missing rate yielded results similar to those based on the complete data. The 60% missing rate resulted in large bias and overestimated standard errors. The 40% missing rate yielded results understandably between 20% and 60%. The performance of multiple imputation under different missing rates in SCED data needs to be further researched. Practical issues relate to implementing multiple imputation are presented in Appendix B.

Multiple imputation has several advantages over *ad hoc* methods (such as mean substitution, listwise deletion) because it (a) retains data already collected, (b) maintains the design structure of a SCED study, (c) avoids potential bias that can result from deleting a participant or a session/interval due to missing data, and (d) captures uncertainty surrounding imputed scores. With the advent of high-speed computers and algorithms, multiple imputation has been increasingly applied by social scientists as a missing data method. It is a statistically proper approach that yields efficient and unbiased estimates for parameters and *SE* under the MAR mechanism (Little & Rubin, 2002). It is applicable for any pattern of missing data and a moderate amount of missing data (e.g., Dong & Peng, 2013). Even though a number of issues surrounding multiple imputation require additional research, we demonstrated the feasibility of applying multiple imputation to treat missing data in a SCED context. Multiple imputation is not making up data. Schafer (1999) illustrated the theoretical framework of multiple imputation, which we highly recommend. For advanced readings, consider Little and Rubin (2002), Rubin (1987), Schafer (1997), SAS Institute Inc. (2015), and Mächler (2015). The latest versions of major statistical software (SAS, SPSS, Stata) and R packages offer multiple imputation capability that makes this missing data method accessible and user-friendly.

The authors of the WWC Handbook did not recognize or recommend any proper missing data method for SCED studies. They preferred analyses be conducted on actual, observed data (WWC, 2017). They encouraged reporting the statistical package that treats missing data, or to adjust *p*-values and standard errors, if necessary, in the presence of missing data. Thus, there is a void in the WWC Handbook on how missing data can be treated properly in order to support the claims made about an intervention effect. According to a recent published checklist on Single-Case Reporting guideline in BEhavioral Interventions, abbreviated as SCRIBE 2016 (Tate et al., 2017), there is a requirement to document sequences completed, as well as the number of trials from each session for each participant, although, the SCRIBE 2016 checklist does not require an explanation of strategies for handing missing data. Because the importance of SCED research in establishing and confirming evidence-based practices has been increasingly affirmed (Horner et al., 2005; Shadish & Sullivan, 2011; Smith, 2012), it is imperative that SCED research be conducted at the highest level of rigor to yield credible and generalizable results. Missing SCED data should not be ignored in visual analysis or in statistical inferences; they should be properly handled. Multiple imputation can significantly improve the inferential validity of single-case studies in applied behavior analyses.

References

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates. In *Proceedings of the thirtieth annual SAS® Users Group International conference* (paper 112-30). SAS Institute Inc., Cary, NC. Retrieved from http://www2.sas.com/proceedings/sugi30/112-30.pdf

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. In *Proceedings of the thirtieth annual SAS® Users Group International conference* (paper 113-30). SAS Institute Inc., Cary, NC. Retrieved from http://www2.sas.com/proceedings/sugi30/113-30.pdf

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469. doi: 10.1111/j.1467-842X.2001.tb00294.x

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. Statistics in Medicine, 26(6), 1368-1382. doi: 10.1002/sim.2619

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159-186). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Carpenter, J., & Goldstein, H. (2004). Multiple imputation using MLwiN. *Multilevel Modelling Newsletter, 16*(2), 9-18. Retrieved from http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/new16-2.pdf

Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification, 39*(6), 835-869. doi: 10.1177/0145445515603706

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi: 10.1037/1082-989x.6.4.330

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation, 78*(1), 69-84. doi: 10.1080/10629360600903866

Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. (1995). *Analysis of longitudinal data*. New York, NY: Oxford University Press.

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus 2*, 222. doi: 10.1186/2193-1801-2-222

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Fahoome, G. (2002). Twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods, 1*(2), 248-268. doi: 10.22237/jmasm/1036110540

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Graham, J. W., Olchowski, A., & Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213. doi: 10.1007/s11121-007-0070-9

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Heitjan, D. F., & Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 40*(1), 13-29. doi: 10.2307/2347902

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165-179. doi: 10.1177/001440290507100203

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*(1), 79-90. doi: 10.1198/000313007X172556

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57*(4), 229-232. doi: 10.1198/0003130032314

International Business Machines Corporation. (2011). IBM SPSS missing values 20 [computer software]. Armonk, NY: International Business Machines Corporation.

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y.-Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions, 8*(2), 88-99. doi: 10.1177/10983007060080020701

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

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. *Psychometrika, 17*(4), 421-428. doi: 10.1007/BF02288917

Mächler, M. (2015). Missing data imputation etc: Literature and R packages. Retrieved from http://stat.ethz.ch/~maechler/adv_topics_compstat/MissingData_Imputation.html

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10*(1), 1-28.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole Publishing Company.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association, 58*(301), 216-230. doi: 10.1080/01621459.1963.10500843

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*(3), 254-265. doi: 10.1007/s10864-012-9153-1

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284–299. doi: 10.1016/j.beth.2010.08.006

Peng, C.-Y. J., & Chen, L.-T. (2015). Algorithms for assessing intervention effects in single-case studies. *Journal of Modern Applied Statistical Methods, 14*(1), 276-307. doi: 10.22237/jmasm/1430454060

Peng, C.-Y. J., Harwell, M., Liou, S.-M., Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 31-78). Charlotte, NC: Information Age Pub.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-96. Retrieved from http://www.statcan.gc.ca/pub/12-001-x/2001001/article/5857-eng.pdf

Rubin, D. B. (1987). *Multiple imputation for nonresponses in survey*. New York, NY: Wiley & Sons. doi: 10.1002/9780470316696

SAS Institute Inc. (2015). *SAS/STAT® 14.1 user's guide*. Cary, NC: SAS Institute Inc. Retrieved from https://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC. doi: 10.1201/9781439821862

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3-15. doi: 10.1177/096228029900800102

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi: 10.1037/1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571. doi: 10.1207/s15327906mbr3304_5

Schenker, N., & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis, 22*(4), 425-446. doi: 10.1016/0167-9473(95)00057-7

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971-980. doi: 10.3758/s13428-011-0111-y

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510-550. doi: 10.1037/a0029312

Smith, J. D., Borckardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time-series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data? *Behavior Therapy, 43*(3), 679-685. doi: 10.1016/j.beth.2011.10.001

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Allyn & Bacon.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Needham Heights, MA: Allyn & Bacon.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H.,…Wilson, B. (2017). The Single-Case Reporting guideline in BEhavioural Interventions 2016 statement. *Neuropsychological Rehabilitation, 27*(1), 1-15. doi: 10.1080/09602011.2016.1190533

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242. doi: 10.1177/0962280206074463

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*(6), 681-694. doi: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

van Buuren, S, Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064. doi: 10.1080/10629360600810434

van de Wiel, M. A., & Di Bucchianico, A. (2001). Fast computation of the exact null distribution of Spearman's ρ and Page's L statistic for samples with and without ties. *Journal of Statistical Planning and Inference, 92*(1-2), 133-145. doi: 10.1016/S0378-3758(00)00166-X

What Works Clearinghouse. (2017). *Standards handbook, version 4.0*. Washington, DC: What Works Clearinghouse. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. doi: 10.1002/sim.4067

Appendix A: SAS Codes for Performing Multiple Imputation on the Lambert Data

Part A

```
*Impute missing meanssr1 + complete ar_ssr1 from DATA=AB_miss  ---------
------------------------;

PROC MI DATA=AB_miss seed=13951639 out=outAB_MI_meanssr1_ar_ssr1;
     EM MAXITER=400;               /* set max. iterations of EM
algorithm to 400     */
     MCMC PRIORS=JEFFREYS;         /* set MCMC’s prior to noninformative
or Jeffreys */
     VAR meanssr1 ar_ssr1;
TITLE 'MI for meanssr1+ar_ssr1'; RUN;

*Impute missing meanrc1 and complete accuracy_rc1 from DATA=AB_miss  ---
------------------------;

PROC MI DATA=AB_miss seed=13951639 out=outAB_MI_meanrc1_accuracy_rc1;
     EM MAXITER=400;
     MCMC PRIORS=JEFFREYS;
     VAR meanrc1 accuracy_rc1;
TITLE 'MI for meanrc1+accuracy_rc1'; RUN;

*Impute missing meanssr2 and complete ar_ssr2 from DATA=AB_miss  -------
------------------------;

PROC MI DATA=AB_miss seed=13951639 out=outAB_MI_meanssr2_ar_ssr2;
     EM MAXITER=400;
     MCMC PRIORS=JEFFREYS;
     VAR meanssr2 ar_ssr2;
TITLE 'MI for meanssr2+ar_ssr2'; RUN;

*Impute missing meanrc2 and complete accuracy_rc2 from DATA=AB_miss  ---
------------------------;
```

```
PROC MI DATA=AB_miss seed=13951639 out=outAB_MI_meanrc2_accuracy_rc2;
     EM MAXITER=400;
     MCMC PRIORS=JEFFREYS;
     VAR meanrc2 accuracy_rc2;
TITLE 'MI for meanrc2+accuracy_rc2'; RUN;
```

Part B

```
* Performing paired t-test for adjacent SSR and RC mean differences-----
------------------------;

PROC TTEST DATA=MI_4means_diff;
VAR diffssr1rc1_mi diffssr2rc2_mi;
BY _imputation_; RUN;
TITLE 'Paired t-test based on MI between one SSR phase and one RC
phase';
RUN;
```

Part C

```
*Performing descriptive analyses of means and SDs of five imputed data
sets---------------------;

PROC MEANS DATA=MI_4means;
VAR meanssr1 meanrc1 meanssr2 meanrc2 varssr1 varrc1 varssr2 varrc2;
BY _imputation_; RUN;
TITLE 'Means and SDs of five imputed data sets';
RUN;
```

Appendix B

When implementing multiple imputation, other practical issues, such as, the selection of auxiliary variables, the specification of an imputation model, the number of imputations, the multivariate normality assumption and rounding imputed values for categorical variables must be considered.

The Selection of Auxiliary Variables

According to Collins, Schafer, and Kam (2001), auxiliary variables are (a) variables that are associated with the missing mechanism, MAR for multiple imputation, and (b) variables that are correlated with the variables with missing data. They are included in an imputation model in order to help generate imputed scores for missing data. For the Lambert data set, we selected four auxiliary variables, one for each imputation model, based on their strong correlation with the missing scores, preferably at least 0.4 in absolute values (Enders, 2010), completeness (van Buuren, Boshuizen, & Knook, 1999), and variability (Enders, 2010; van Buuren et al., 1999). If missing data are to be expected in a SCED study, it is desirable to collect information about potential auxiliary variables such as, age (a demographic variable), classroom (a setting variable), or date or session number (a time-related variable). An auxiliary variable can be continuous, categorical/nominal, or ordinate-level variables (Collins et al., 2001). Even when a participant misses a session, the information about his/her auxiliary variable can still be available or collected. The inclusion of auxiliary variable(s) is beneficial to the success of multiple imputation, especially when the imputation model is simple like the four models specified for Lambert data.

The Specification of an Imputation Model

Multiple imputation requires two models: the imputation model used in Step 1 and the analysis model used in Step 2. Theoretically, multiple imputation assumes that the two models are the same. In practice, they can be different (Schafer, 1997). An appropriate imputation model is the key to the effectiveness of multiple imputation; it should have the following two properties. First, an imputation model should include useful variables. Schafer (1997) and van Buuren et al. (1999) recommended three kinds of variables to be included in an imputation model: (1) variables that are of theoretical interest, and auxiliary variables, namely, (2) variables that are associated with the missing mechanism, or (3) variables that are correlated with the variables with missing data. The first kind of variables is necessary, because omitting them will diminish the relationship between these variables and other variables in the imputation model. The second kind of variables makes the MAR assumption more plausible, because they account for the missing mechanism. The third kind of variables helps to estimate missing values more precisely. Thus, each kind of variables has a unique contribution to the multiple imputation process. However, including too many variables in an imputation model may inflate the variance of estimates, or lead to non-convergence. Thus, researchers should carefully select variables to be included into an imputation model.

An imputation model should be general enough to capture the assumed structure of the data. If an imputation model is more restrictive, namely, making additional restrictions, than an analysis model, one of two consequences may follow. One consequence is that the results are valid but the conclusions may be conservative (i.e., failing to reject the false null hypothesis), if the additional restrictions are true (Schafer, 1999). The second consequence is that the results are invalid because one or more of the restrictions is false (Schafer, 1999). For example, a restriction may restrict the relationship between a variable and other variables in the imputation model to be merely pairwise. Consequently, any interaction effect will be biased toward zero. To handle interactions properly in multiple imputation, Enders (2010) suggested that the imputation model include the product of the two variables if both are continuous. For categorical variables, Enders suggested performing multiple imputation separately for each subgroup defined by the combination of the levels of the categorical variables.

For the Lambert data set, we constructed a simple imputation model that consisted of one auxiliary variable with complete data and one variable with missing data to ensure the convergence of the MCMC method. Because the baseline phase and the intervention phase had different numbers of sessions in Classes A and B, we decided to impute the missing mean of each phase, instead of missing session scores. Constructing a simple imputation model based on a phase mean may prove to be a necessary strategy in most SCED studies due to its characteristically small n and differing number of sessions per phase.

The Number of Imputations

The number of imputations necessary in multiple imputation is a function of the rate of missing information in a data set. A data set with a large amount of missing information requires more imputations. Rubin (1987) provided a formula to compute the relative efficiency (RE) of imputing m times, instead of an infinite number of times. However, methodologists have not agreed on the optimal number of imputations. Schafer and Olsen (1998) suggested that "in many applications, just 3-5 imputations are sufficient to obtain excellent results" (p. 548). Schafer and Graham (2002) were more conservative in asserting that 20 imputations were enough in many practical applications to remove noises from estimations. Graham, Olchowski, and Gilreath (2007) commented that RE should not be an important criterion when specifying m, because RE has little practical meaning. Other factors, such as, the *SE*, *p*-value, and statistical power, are more related to empirical research and should also be considered, in addition to RE. Graham et al. reported that statistical power decreased much faster than RE, as the rate of missing information increases and/or *m* decreases. White et al. (2011) suggested that the number of imputations should be greater than or equal to the percentage of missing observations in order to ensure an adequate level of reproducibility.

Because SCED data sets usually do not contain a large number of participants, phases, or sessions, nor will a complex or large imputation model be applied, we recommend that researchers and analysts start with $m = 5$ imputations to ensure that the imputation process converges and stabilizes. PROC MI defaults *m* to 5. In general, it is a good practice to specify a sufficient *m* to ensure the convergence of multiple imputation within a reasonable computation time (Dong & Peng, 2013).

The Multivariate Normality Assumption

The MCMC method implemented in SAS, R, and other statistical packages (e.g., Stata) assume multivariate normality for variables included in the imputation model. It has been shown that multiple imputation based on the multivariate normal assumption can provide valid estimates even when this assumption is violated (Demirtas, Freels, & Yucel, 2008; Schafer, 1997, 1999). Furthermore, this assumption is robust when the sample size is large and when the missing rate is low, although the definition for a large sample size or for a low rate of missing is not specified in the literature for between- or within-subject designs (Schafer, 1997). When an imputation model contains categorical variables, one cannot use the MCMC or regression method directly. If the missing data pattern is arbitrary, MCMC based on other probability models (such as the joint distribution of normal and binary) can be used for imputation. If the missing data pattern is monotonic or univariate, logistic regression and discriminant function analysis can substitute for the regression method. SAS, R, Stata, or SPSS provide a wide range of options for implementing multiple imputation. Interested readers are referred to volume 45 of the *Journal of Statistical Software* (http://www.jstatsoft.org/issue/view/v045) for software developments on multiple imputation up to 2011.

Rounding Imputed Values for Categorical Variables

A common practice is to round the imputed value to the nearest integer, or to the nearest plausible value. However, this intuitive strategy could not provide desirable results for binary missing values (Ake, 2005; Allison, 2005; Bernaards, Belin, & Schafer, 2007; Enders, 2010). Horton, Lipsitz, and Parzen (2003) showed analytically that rounding the imputed values of a binary variable led to biased estimates, whereas imputed values without rounding led to unbiased results. Unfortunately, even less is known about the effect of rounding on imputed values of ordinal variables with three or more levels. It is possible that as the level of the categorical variable increases, the effect of rounding decreases. Several factors influence the performance of a rounding strategy, such as, the missing mechanism, the size of the model, distributions of the categorical variables. These factors are not within a researcher's control. Additional research is needed to explore and identify viable strategies for dealing with the rounding issue for categorical variables during multiple imputation for missing SCED data.