

Experimental Determination of Effectiveness of an Individual Information-Centered Approach in Recovering Step-Count Missing Data

Minsoo Kang

*Department of Health and Human Performance
Middle Tennessee State University*

Weimo Zhu

*Department of Kinesiology and Community Health
University of Illinois at Urbana-Champaign*

Catrine Tudor-Locke

*Department of Exercise and Wellness
Arizona State University East*

Barbara Ainsworth

*Department of Exercise and Nutritional Sciences
San Diego State University*

Missing values are a common phenomenon in physical activity research, which has a negative impact on the quality of the data collected. The purpose of this study was to determine empirically the effectiveness of an individual information-centered (II-centered) approach in recovering step-count missing values by comparing the performance of the II-centered approach with the traditional group information-centered approach. Using pedometers, the step counts of 117 participants were measured for 21 consecutive days. A semisimulation approach was used to create a missing data set and several conditions of weekdays, weekend days, or both, were compared under each approach. Two accuracy indexes, Root Mean Square Difference and Mean

Signed Difference, were used to determine the effectiveness of the recovery methods, and paired *t* tests were used to examine the mean differences between the original and the replacement values. The II-centered approach produced a more accurate recovery method. The impact of the findings and future research directions in recovering physical activity data are discussed.

Key words: physical activity, missing value, recovery method, group information-centered approach, individual information-centered approach

Physical activity plays an important role in the maintenance of health and effective functioning in people of all ages. Regular physical activity lowers the risk of premature mortality (Blair et al., 1995), reduces cardiovascular disease (Blair & Connelly, 1996; Pate et al., 1995), decreases depression and anxiety (Landers & Petruzzello, 1994; Morgan, 1994), and improves health-related quality of life by enhancing psychological well-being (McAuley & Rudolph, 1995). Government agencies and major national public organizations, such as the Centers for Disease Control and Prevention and the American Heart Association, have all given official statements and reports to encourage people to participate in regular physical activity. The 1996 U.S. Surgeon General's benchmark report concluded that significant health benefits can be obtained by including 30 min of moderately intense physical activity (e.g., brisk walking) on most or all days of the week (U. S. Department of Health and Human Services, 1996).

The recognition of the importance of physical activity has led to greater interest in assessing physical activity behavior. Many surveys or instruments (e.g., physical activity questionnaires and motion sensors) have been developed to measure physical activity behavior (Dale, Welk, & Matthews, 2002). The pedometer is becoming one of the most commonly used practical measures to assess physical activity in exercise and physical activity interventions (Tudor-Locke & Bassett, 2004; Tudor-Locke, Williams, Reis, & Pluto, 2002; Welk et al., 2000). Pedometers are inexpensive, simple movement counters that can estimate physical activity over a long period of time. The measurement is based on tracking the number of steps a person has walked, that is, step counts (Bassett & Strath, 2002).

One of the challenges in measuring step counts, like other types of physical activity, is its high intra-individual variability, that is, due to many factors such as time of year, weather, type of day, and places, individuals often participate in physical activity at various levels during a time period. Because of the high intra-individual variability, researchers have tried to determine more appropriate data collection periods for physical activity research. Gretebeck and Montoye (1992) reported that at least 5 or 6 consecutive days are needed to minimize the intra-individual variability to an acceptable degree. They also recommended that both weekdays and weekend days should be included. Baranowski and de Moor (2000) noted

that between 6 and 8 days are needed to obtain a stable estimation of children's physical activity participation. Because a stable measure of physical activity often requires data collection on multiple days, missing values have become a common phenomenon. This is also true for step counts measured by a pedometer.

Missing data can potentially threaten the validity of a physical activity research study. First, incomplete data represent an absence of information, which might mean that the researcher has lost important information from the data. Second, statistical power is decreased with the loss of large amounts of data when the cases with the missing values are deleted from the data set. Finally, many statistical procedures assume complete cases, and missing values may violate the assumption. Missing data have a negative impact on the quality of research studies and have forced researchers to spend more time and money in recruiting larger sample sizes in hopes of having enough completed cases to properly analyze their data (Statistical Package for the Social Sciences [SPSS], 1997). To eliminate or reduce the threat of missing data, statistical methods have been developed to try to recover the missing values.

Statistical missing data methods can be divided into two categories: deletion and imputation. In the deletion approach the cases with missing values are simply deleted. List-wise and pair-wise deletion methods are two commonly used methods to deal with missing data (Allison, 2001; Little & Rubin, 1987). Because the computation process is relatively quick and simple, those deletion methods are the default options for many statistical packages, such as SPSS and SAS. Deletion methods, however, may not be the best choice because a large amount of data could be lost. The deletion decreases the sample size, increases standard errors and, therefore, reduces statistical power (Acock, 1997; Roth, 1994; Stolzenberg & D'Alessio, 1993).

The imputation approach replaces missing values with new estimates. Many recovery methods have been developed under this approach. Mean substitution, regression imputation, and expectation maximization (EM) approaches are commonly used (Acock, 1997; Allison, 2001; Little & Rubin, 1987). Other missing recovery methods are also available, for example, cold-deck imputation, hot-deck imputation, dummy-variable adjustment, structure equation modeling approach, and multiple imputation (MI). For more information on these methods, the reader may refer to Acock (1997); Allison (2001); Dempster, Laird, and Rubin (1977); Hox (1999); Little and Rubin (1987); and Raymond and Roberts (1987).

Most prior research that has applied recovery methods used only group information (GI) where an individual's missing value is replaced by a summary (e.g., mean) from the group to which the individual belongs. The GI-centered approach, however, may not be appropriate in handling step-count data because: (a) multiple days are often needed to collect the step-count data due to the high intra-individual variability nature of physical activity behaviors, for example, the amount of physical activity a person performs on a Monday could differ greatly from that of a

Sunday; and (b) when repeated measures are employed, replacements based on intra-individual information should be able to generate a more accurate recovery of the missing values. Laird (1988) pointed out that when data are collected repeatedly over time on the same experimental units, for example, individuals, the use of the GI-centered approach is inappropriate, results in a loss of efficiency, and may bias the results. Schafer and Graham (2002) recommended procedures that use all available data for each participant. They believed that missing information could be partially recovered from the earlier or later data from the same individuals. Because the missing information is generated from the same individuals, replacing missing values based on the individual information-centered (II-centered) approach should be able to create a better recovery of the missing values.

Conceptually, the II-centered approach is based on the data from each individual, that is, replacing a missing value using a summary from the rest of the data from the same individual. Technically, II-centered approach can be easily implemented using existing statistical analysis software by transposing the data set from the variables located in columns and rows. Transposing the data set for the II-centered approach is illustrated in a step-by-step manner in the Appendix. No empirical evidence, however, is available to confirm that the II-centered approach can recover missing values more accurately than the GI-centered approach.

The nature of the data, or characteristics, must also be taken into consideration when applying the II-centered approach to physical activity data. This is because the type of days in physical activity measurement may have its unique pattern, for example, weekdays' physical activity may differ from weekend days' physical activity (Matthews, Ainsworth, Thompson, & Bassett, 2002). When applying a missing-value recovery method, therefore, one must determine which combination of day(s) is to be used for generating a replacement for a missing value, for example, use the information of weekdays only, weekend days only, or both. The impact of the combination should be examined empirically. Finally, the impact of the number of missing values on the recovery should also be examined. Roth (1994) addressed the amount of missing data as the key factor which should be considered when choosing missing value recovery methods. Roth hypothesized that the accuracy of recovering decreases as the number of missing values increases.

The effectiveness of a missing-value recovery method is often determined by comparing the degree of recovery from the methods. Many previous studies compared recovery methods using either simulation data or an available large data set (Acock, 1997). Those kinds of data sets, however, are often far from real data and the findings on a recovery method may not apply to the unique pattern of a middle- or small-size, real-life data set. As a result, missing values may not be well recovered. This study introduced a semisimulation approach in which the comparisons among the recovery methods are based on the data set to which the methods are going to apply. More specifically, a set of missing values was created from a *nonmissing* sample from the original data set, and the utility of different recovery methods

was evaluated based on the semisimulated missing data set. Because the true values of the missing values, deleted on purpose, were known, the accuracy of each method could be compared.

Using the semisimulation approach, the purposes of this study were threefold: (a) determine the impact of various combinations of weekdays, weekend days, or both, in recovering missing values using the II-centered approach; (b) compare the recovery utility and effectiveness of the II-centered approach with the GI-centered approach; and (c) examine the impact of the amount of missing data on the accuracy of data replacement.

METHOD

Participants and Data Collection

A total of 117 participants, aged from 17 to 79 years old, were involved in the study. They were recruited by word of mouth and from posted announcements in the University of South Carolina and Columbia communities. Participants were instructed to wear the pedometer during waking hours and recorded their step counts for 21 consecutive days with a random starting date (i.e., some started the data collection on Mondays and some on the other days). Fifty-four of the 117 participants had complete data, and the remaining 63 had at least one missing data point. The data set used in this study has been previously described by Tudor-Locke et al. (2003). All participants read and signed a written informed consent form approved by the University of South Carolina Institutional Review Board.

Semisimulation Data Generation

A semisimulation design was employed to create a data set to evaluate the recovery methods. Fifty-four of 117 participants, who had no missing values, were first selected from the data to form a *nonmissing* sample. Then, a *missing* sample, also with 54 participants, was randomly selected from the remaining data. Based on the missing data pattern in the *missing* sample, a new set of semisimulation data was created in a one-to-one fashion using the data of the *nonmissing* sample. If, for example, the first person in the *missing* sample missed the data collection on the first Wednesday and the second Thursday, the data on those days for the first participant in the *nonmissing* sample would be removed. This process was repeated until the 54th person's missing data pattern in the *missing* sample was duplicated in the 54th person in the *nonmissing* sample. Because the true values of intentionally missing data in the *nonmissing* sample were known, these data enabled us to judge which recovery method can best recover the true values of the missing data (see Figure 1).

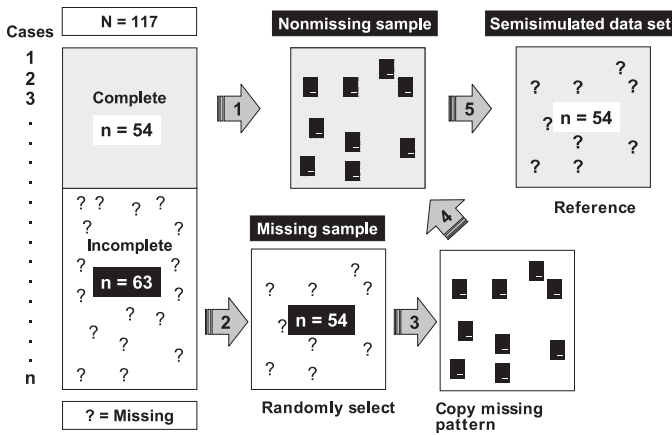


FIGURE 1 The procedure to create a semisimulated data set.

Recovery Method and Related “Which Day?” Conditions

Mean substitution was employed in the study and applied to several “which day?” conditions to evaluate the best recovery of missing values. The mean substitution method uses the group’s average on a variable, that is, the mean of other participants who had no missing values on the same variable. This method was selected because it is included in most statistical packages and easy to apply. Replacements from this method can be easily implemented by using SPSS software. To answer the “which day?” question, several conditions were examined, which were combinations of weekdays and weekend days information depending on the type of a missing day. Four conditions were calculated by the II-centered approach:

1. The remaining days; for example, to replace a missing Monday step count by the mean of step counts in the remaining days of the same week, along with the days in the other 2 weeks.
2. The remaining weekdays or weekend days, depending on the type of a missing day; for example, to replace a missing Monday by the mean of Tuesday to Friday of the same week, and Mondays to Fridays in the other 2 weeks.
3. The remaining weekdays or weekend days in the same week when the missing value occurred; for example, to replace a missing Tuesday of the first week by the mean of Monday, and Wednesday to Friday in the same, first week.
4. The same days, but in other weeks; for example, to replace a missing Monday by the mean of the Mondays in the other 2 weeks.

If the missing values occurred in 2 or more days, the same principle was also applied for those conditions. For example, in Condition 3, if Monday and Wednesday were missing, they were replaced by the mean of Tuesday, Thursday, and Friday of the same week. If an individual's missing values occurred on Saturday and Sunday, however, replacements could not be made for Condition 3 because of the lack of references. In that case, the individual was not included in the analysis.

For comparison, four conditions of weekdays and weekend days were also employed by the GI-centered approach:

5. The day the missing value occurred; for example, to replace a missing Monday by the group's average for the Monday of the same week.
6. The weekdays or weekend days depending on the type of a missing day; for example, to replace a missing Monday by the group's average for all weekdays, Monday through Friday.
7. The weekdays or weekend days in the same week when the missing value occurred; for example, to replace a missing Monday of the second week by the group's average for the second week, Monday through Friday.
8. The same days when the missing value occurred; for example, to replace a missing Saturday by the group's average for all Saturdays, over the 3 weeks.

Condition 5 is the traditional mean substitution method, and Conditions 6 to 8 are extensions of Condition 5 to match the multiple day conditions employed in the II-centered approach.

Data Analyses

The impact of various combinations of weekdays and weekend days in recovering missing values was evaluated by comparing the original, known values purposely removed from the semisimulated missing data set with the replacements estimated based on the different recovery conditions. Two indexes, Root Mean Square Difference (RMSD) and Mean Signed Difference (MSD), were used to determine the effectiveness of the various recovery conditions. RMSD was calculated by the differences between the original and replacement values, which were then squared, averaged, and square-rooted. The formula for RMSD is as follows:

$$RMSD = \sqrt{\frac{\sum_{j=1}^N (\text{original value}_j - \text{replacement value}_j)^2}{N}}, \quad (1)$$

where N refers to the total number of missing data points. The RMSD index provides the degree of bias that may exist with replacing the missing values. A smaller RMSD represents a more accurate recovery of missing values. To calculate MSD, the differences between the original and replacement values were averaged. The formula for MSD is as follows:

$$MSD = \sqrt{\frac{\sum_{j=1}^N (\text{original value}_j - \text{replacement value}_j)}{N}}. \quad (2)$$

The MSD index presents the direction of bias that may be caused by the missing data recovery. A close-to-zero MSD represents a more accurate recovery of missing values.

Paired t tests were used to examine the mean differences between the original and the replacement values. The alpha level was set at .006, adjusted by the Bonferroni technique (i.e., .05/8). In addition, the RMSD and MSD indexes for the II-centered approach were compared by the number of missing values (i.e., only one missing value and two or more missing values) to determine the impact of the amount of missing data on the recovery. These analyses were completed using Microsoft EXCEL and SPSS 11.0 statistical software (SPSS Inc., 2001).

RESULTS

Descriptive Statistics

In the original data set ($N = 117$), 46% ($n = 54$) of the cases had no missing values, 24% ($n = 28$) of the cases had only one, 16% ($n = 19$) of the cases had two to four, and 14% ($n = 16$) had five or more missing values. The highest mean steps were found on Wednesdays ($M \pm SD: 7,988.69 \pm 4,283.04$) and Sundays' mean steps were reported as the lowest ($6,226.41 \pm 3,973.32$). Grand mean steps were 7,561.89. The mean and standard deviation and minimum and maximum values of the step count data are presented by day and different data set in Table 1.

The percentage of total missing values in the semisimulated missing data set ($n = 54$) was 10.14%. Similar to the original data set, the highest mean steps were found on Wednesdays ($8,667.66 \pm 4,059.68$), and Sundays' mean steps were reported as the lowest ($6,686.07 \pm 3,606.11$; see Table 1). Grand mean steps were 7,946.40. Sundays had the highest number of missing values, and Thursdays had the least, as follows: Sundays ($n = 21$), Wednesdays ($n = 19$), Tuesdays ($n = 18$), Fridays ($n = 17$), Saturdays ($n = 16$), Mondays ($n = 14$), and Thursdays ($n = 10$). There was little variation in the number of missing values for 3 weeks: Week 1 ($n = 36$), Week 2 ($n = 40$), and Week 3 ($n = 39$).

TABLE 1
 Statistical Summary of Step-Count Information by Day and Data Set

Days	Original Data				Artificial Data			
	<i>M</i>	<i>SD</i>	<i>Minimum</i>	<i>Maximum</i>	<i>M</i>	<i>SD</i>	<i>Minimum</i>	<i>Maximum</i>
Mondays	7,676.87	4,278.29	1,062.82	2,4107.54	8,312.78	3,939.72	1,253.76	1,9794.68
Tuesdays	7,826.25	4,442.34	1,042.85	2,5997.09	7,736.91	3,595.72	1,398.75	2,1932.38
Wednesdays	7,988.69	4,283.04	1,073.90	2,6999.67	8,667.66	4,059.68	2,391.72	2,0024.96
Thursdays	7,521.42	3,958.34	1,046.88	2,5592.35	7,718.51	3,344.62	1,760.68	1,9258.40
Fridays	7,938.87	4,296.56	1,025.81	2,1847.90	8,575.46	4,055.56	1,025.81	1,9527.04
Saturdays	7,754.70	4,443.79	1,002.89	2,5603.60	7,927.44	3,961.69	1,450.76	2,0261.92
Sundays	6,226.41	3,973.32	1,023.98	2,0454.47	6,686.07	3,606.11	1,079.81	1,7963.93
Total	7,561.89	4,239.38			7,946.40	3,794.73		

Note. Original data = data from original data set, $N = 117$; Artificial data = data from semisimulated missing data set, $n = 54$.

TABLE 2
 Root Mean Square Difference by Method

Missing Recovery Methods	Conditions			
	1	2	3	4
II-centered approach	3,020.53	2,793.92	3,483.51	3,092.70
	5	6	7	8
GI-centered approach	3,822.43	3,856.84	3,864.08	3,823.56

Note. See the Method section on the description of the conditions.

Missing Recovery

The results of RMSD for both II-centered and GI-centered approaches are summarized in Table 2. A smaller RMSD value represents a better recovery of the missing values. The RMSD of the II-centered and GI-centered conditions ranged from 2,793.92 to 3,483.51 and from 3,822.43 to 3,864.08, respectively. Overall, the II-centered approach showed smaller RMSDs than the GI-centered approach. For the II-centered approach, Condition 2, that is, replace a missing value by the mean of the remaining weekdays or weekend days, depending on the type of a missing day, was most effective in recovering the missing values (RMSD = 2,793.92). For the GI-centered approach, Condition 5, that is, replace a missing value by the group's average of the day when the missing value occurred, was most effective in recovering the missing values (RMSD = 3,822.43).

The results of MSD for both II-centered and GI-centered conditions are summarized in Table 3. MSD was included to show the direction and the degree of bias

TABLE 3
Mean Signed Difference by Method

<i>Missing Recovery Methods</i>	<i>Conditions</i>			
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
II-centered approach	-802.23	-610.83	-683.07	-587.40
<i>t</i> values	-2.94*	-2.39	-1.95	-2.01
	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>
GI-centered approach	-1,074.97	-1,085.81	-1,083.97	-1,089.22
<i>t</i> values	-3.13*	-3.13*	-3.12*	-3.17*

Note. See the method section on the description of the conditions.

* $p < .006$ (adjusted by the Bonferroni technique).

that may be caused by missing data recovery methods. The smallest MSD values were found from the II-centered approach (ranging from -587.40 to -802.23), which represents a more accurate recovery. Highest MSD indexes were found in the GI-centered approach (ranging from -1,074.97 to -1,089.22). Negative MSDs across all conditions indicate that predicted missing values tended to be overestimated. Paired *t*-test results are presented under each condition in Table 3. After adjusting the alpha level by the Bonferroni technique, no statistically significant mean differences were found between the original and the replacement values in Conditions 2, 3, and 4 for the II-centered approach. All conditions in the GI-centered approach showed statistically significant mean differences.

The RMSD and MSD results for the II-centered approach by the number of missing values are summarized in Table 4. RMSD indexes were the smallest across all conditions when there was only one missing value. When the missing value was replaced by the mean of the remaining weekdays or weekend days, depending on the type of a missing day, that is, Condition 2, the effectiveness in recovering missing values was high on both only one missing value and two or more missing values. Similar to the finding from RMSD indexes, the smallest MSD indexes were found where there is only one missing value. Negative MSD indexes across all conditions implied an overestimation of predicted missing values.

DISCUSSION

As expected, the II-centered approach produced smaller RMSDs and MSDs than the GI-centered approach, indicating that the II-centered method might be more efficient in handling missing values in physical activity research. These findings were comparable to the findings by Albridge and others (Albridge, Standish, &

TABLE 4
RMSD and MSD by the Number of Missing Values

<i>Recovery Indices</i>	<i>Conditions</i>			
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
RMSD				
Overall	3,020.53	2,793.92	3,483.51	3,092.70
Only one missing	2,297.62	2,151.65	2,803.43	2,845.51
Two or more missing	3,218.84	2,971.24	3,675.72	3,168.15
MSD				
Overall	-802.23	-610.83	-683.07	-587.40
Only one missing	-476.43	-495.31	-208.83	-476.72
Two or more missing	-907.08	-648.01	-855.53	-624.29
Number of participants				
Overall	54	54	50	53
Only one missing	28	28	28	28
Two or more missing	26	26	22	25

Note. RMSD = Root Mean Square Difference; MSD = Mean Signed Difference. See the Method section on the description of the conditions.

Fries, 1988). They applied several time-oriented techniques (e.g., repeating the previous values), similar to the II-centered approach, into a time-oriented data bank, the American Rheumatism Association Medical Information System collected from 1969 to 1986. They found that the replacements from time-oriented techniques were better predictors than those from the traditional standard methods, such as regression (Albridge et al., 1988).

The data from physical activity measurement have a unique pattern that needs to be taken into account when replacing missing values. This study takes into consideration this uniqueness by comparing several conditions or combinations of weekdays, weekend days, or both, to generate a better replacement for a missing value. The impact of those combinations was examined empirically. Among the methods of the II-centered approach examined, replacing a missing value by the mean of the remaining weekdays or weekend days depending on the type of a missing day, Condition 2, was found to be the most effective in recovering missing values in the data set. The data collection period from this study was 3 weeks, which allows us to generate unique conditions such as Condition 2 and 4 to find a better replacement for a missing value. Some of the conditions used in this study, however, may not be applied directly to the data set when its collection period is only a week or less. Several conditions and combinations of weekdays and weekend days should be considered based on the data set used to make a better decision on selecting recovery conditions.

One of the unique aspects of this study was to employ a semisimulation approach to modify an existing data set. Because the data characteristics are often not

taken into consideration, traditional ways to simulate data or employ a large existing data set to study the effectiveness of a missing-value recovery method may not be useful for the missing-value recovery practice in middle- or small-size data sets. The semisimulation approach introduced in this study, in contrast, took the data characteristics into consideration in determining the most effective recovery method. Because the true values of the missing data points were known, the accuracy of each method could be compared and thus, a more appropriate method could be identified. To help the reader understand the details of the semisimulation design, a step-by-step instruction on how to replace the missing values is enclosed in the Appendix. The process consists of six steps, and each step includes a short written description and illustration.

Although the findings of this study support the usage of II-centered approach to recover missing values in middle- or small-size physical activity data sets, the reader should be aware of the potential impact of the small sample size. This study included 117 participants, 54 of whom had no missing values. A 54 *nonmissing* sample may be enough to make a decision about which method can best recover the true values of the missing data; however, if the sample size is very small (e.g., a total of 20 participants with 5 having no missing values), it may not be possible to create a representative sample for the empirical procedure described in this study. Future research may focus on ways to create a reliable data set with only a small sample size available.

Because a rather large RMSD was observed, the reader should also be cautious about inferences drawn from a data set with a high percentage of replacements, even if they were derived based on an II-centered approach. A large RMSD means that the missing data may not be accurately recovered. The replacements from different missing data recovery conditions were compared to the original values purposely removed from the semisimulated data set to determine the amount of error. The relatively large RMSD values might be due to the larger variability in people's walking behaviors. The average step counts per day was approximately 8,000 steps, but with a very large standard deviation (about 4,000 steps). Even if the recovery is based on a person's own walking data, the accuracy of the recovery is relatively low because of the large variation in a person's day-to-day walking behavior.

The RMSD has been widely used to evaluate effective missing data recovery methods (Huisman, 2000; Switzer, Roth, & Switzer, 1998) and absolute evaluation criteria have been developed in certain areas. In biology, for example, the RMSD index has been used to find the structural alignment between atoms, and the RMSD criteria are currently available in a measurement unit of Angstroms (i.e., zero = identical structures, between one and three = similar structures, and over 3 = distant structures; Stark, Sunyaev, & Russell, 2003). No such criterion, however, has been created for physical activity data. Because RMSD values are impacted by different units (e.g., steps count by pedometers vs. activity counts by accelerometers)

and sample sizes, the RMSD generated from a particular study is difficult to compare with the RMSD computed from another study. There is thus a need to develop absolute RMSD criteria for physical activity data.

Finally, it should be pointed out that only mean substitution was employed in the current study. The effectiveness of other methods should be explored under the same semisimulation framework. For example, the MI method proposed by Rubin (1987) may provide a useful strategy for dealing with missing values. Some researchers found a drawback imputing one value for each missing value; standard errors are usually small because single imputation methods cannot represent any uncertainty that occurs when the reasons for the missing data are not known (Little & Rubin, 1989). To overcome this limitation of filling in a single value for each missing value, MI creates multiple imputed data sets in which each data set has different imputed values for the missing data. The statistics from each of the data sets are combined in a summary set of findings including overall estimates and standard errors. As the data analysis process has become simplified by the development of MI software including NORM (Schafer, 1997) and SOLAS (Statistical Solutions, 1998), MI has become widely available. Other related methods (e.g., regression imputation and EM) and factors (e.g., the nature of missing data pattern and the amount of missing data) also should be examined in future research.

CONCLUSION

Examination of effective recovery methods for missing values, which allow physical activity researchers to conserve time and effort by preserving data, helps better explain the pattern of physical activity data, and increases the quality of data. Both II-centered and GI-centered approaches were compared in recovering missing values in a step-count data set. In conclusion, the II-centered approach is better in recovering the missing values in step-count data. Among the II-centered methods examined, replacing missing values by the mean of the remaining weekdays or weekend days, depending on the type of a missing day, that is, Condition 2 is the most effective in recovering the missing values in the data set. The information based on missing values, however, should be used cautiously, considering the rather large RMSD values found in this study. Other alternatives that may decrease RMSD (e.g., increase the number of repeated measures or using other statistical models) should be explored in future studies.

ACKNOWLEDGMENT

The data of this study came from a project, which is funded by a supplement to CDC SIP4-99; U48/CCU409664-06 and directed by Dr. Barbara Ainsworth.

REFERENCES

- Acock, A. C. (1997). Working with missing values. *Family Science Review*, *10*, 76–102.
- Albridge, K. M., Standish, J., & Fries, J. F. (1988). Hierarchical time-oriented approaches to missing data inference. *Computers and Biomedical Research*, *21*, 349–366.
- Allison, P. D. (2001). *Missing data*. (Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07–136). Thousand Oaks, CA: Sage.
- Baranowski, T., & de Moor, C. (2000). How many days was that? Intra-individual variability and physical activity assessment. *Research Quarterly for Exercise and Sport*, *71*(Suppl 2), 74–78.
- Bassett, D. R., Jr., & Strath, S. J. (2002). Use of pedometers to assess physical activity. In G. J. Welk (Ed.), *Physical activity assessments for health-related research* (pp. 163–177). Champaign, IL: Human Kinetics.
- Blair, S. N., & Connelly, J. C. (1996). How much physical activity should we do? The case for moderate amounts and intensities of physical activity. *Research Quarterly for Exercise and Sport*, *67*, 193–205.
- Blair, S. N., Kohl, H. W., III, Barlow, C. E., Paffenbarger, R. S., Jr., Gibbons, L. W., & Macera, C. A. (1995). Changes in physical fitness and all-cause mortality: A prospective study of healthy and unhealthy men. *Journal of the American Medical Association*, *273*, 1093–1098.
- Dale, D., Welk, G. J., & Matthews, C. E. (2002). Methods for assessing physical activity and challenges for research. In G. J. Welk (Ed.), *Physical activity assessments for health-related research* (pp. 19–34). Champaign, IL: Human Kinetics.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *B39*, 1–38.
- Gretebeck, R., & Montoye, H. (1992). Variability of some objective measures of physical activity. *Medicine & Science in Sports & Exercise*, *24*, 1167–1172.
- Hox, J. J. (1999). A review of current software for handling missing data. *Kwantitatieve Methoden*, *62*, 123–138.
- Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity*, *34*, 331–351.
- Laird, N. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, *7*, 305–315.
- Landers, D. M., & Petruzzello, S. J. (1994). Physical activity, fitness, and anxiety. In C. Bouchard, R. J. Shephard, & T. Stephens (Eds.), *Physical activity, fitness, and health* (pp. 868–882). Champaign, IL: Human Kinetics.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, *18*, 292–326.
- Matthews, C. E., Ainsworth, B. E., Thompson, R. W., & Bassett, D. R., Jr. (2002). Sources of variance in daily physical activity levels as measured by an accelerometer. *Medicine & Science in Sports & Exercise*, *34*, 1376–1381.
- McAuley, E., & Rudolph, D. (1995). Physical activity, aging, and psychological well-being. *Journal of Aging and Physical Activity*, *3*, 67–96.
- Morgan, W. P. (1994). Physical activity, fitness, and depression. In C. Bouchard, R. J. Shephard, & T. Stephens (Eds.), *Physical activity, fitness, and health* (pp. 851–867). Champaign, IL: Human Kinetics.
- Pate, R. R., Pratt, M., Blair, S. N., Haskell, W. L., Macera, C. A., Bouchard, C., et al. (1995). Physical activity and public health: A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *Journal of the American Medical Association*, *273*, 402–407.
- Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, *47*, 13–26.

- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*, 537–560.
- Rubin, D. B. (1987). *Multiple imputations for nonresponse in surveys*. New York: Wiley.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods, 7*, 147–177.
- Stark, A., Sunyaev, S., & Russell, R. B. (2003). A model for statistical significance of local similarities in structure. *Journal of Molecular Biology, 326*, 1307–1316.
- Statistical Package for the Social Sciences, Inc. (1997). *Missing data: The hidden problem* [SPSS white paper; online]. Retrieved September 2, 2003, from <http://users.bigpond.net.au/david-devaus/Methods-resources/MISSINGP.pdf>
- Statistical Package for the Social Sciences, Inc. (2001). *SPSS 11.0 for Windows*. Chicago: Prentice Hall.
- Statistical Solutions. (1998). *SOLAS for missing data analysis* (Version 1): User reference. Cork, Ireland: Author.
- Stolzenberg, L., & D'Alessio, S. J. (1993). Handling missing data with SPSS/PC+: A test and a tutorial. *Social Science Computer Review, 11*, 185–196.
- Switzer, F. S., III, Roth, P. L., & Switzer, D. M. (1998). Systematic data loss in HRM settings: A Monte Carlo analysis. *Journal of Management, 24*, 763–779.
- Tudor-Locke, C., Ainsworth, B. E., Whitt, M. C., Thompson, R. W., Addy, C. L., & Jones, D. A. (2003). Ambulatory activity and simple cardiorespiratory parameters at rest and submaximal exercise. *Canadian Journal of Applied Physiology, 28*, 699–709.
- Tudor-Locke, C., & Bassett, D. R., Jr. (2004). How many steps/day are enough? Preliminary pedometer indices for public health. *Sports Medicine, 34*, 1–8.
- Tudor-Locke, C., Williams, J. E., Reis, J. P., & Pluto, D. (2002). Utility of pedometers for assessing physical activity. *Sports Medicine, 32*, 795–808.
- U. S. Department of Health and Human Services. (1996). *Physical activity and health: A report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion.
- Welk, G. J., Differding, J. A., Thompson, R. W., Blair, S. N., Dziura, J., & Hart, P. (2000). The utility of the Digi-Walker step counter to assess daily physical activity patterns. *Medicine & Science in Sports & Exercise, 32*(Suppl. 9), S481–S488.

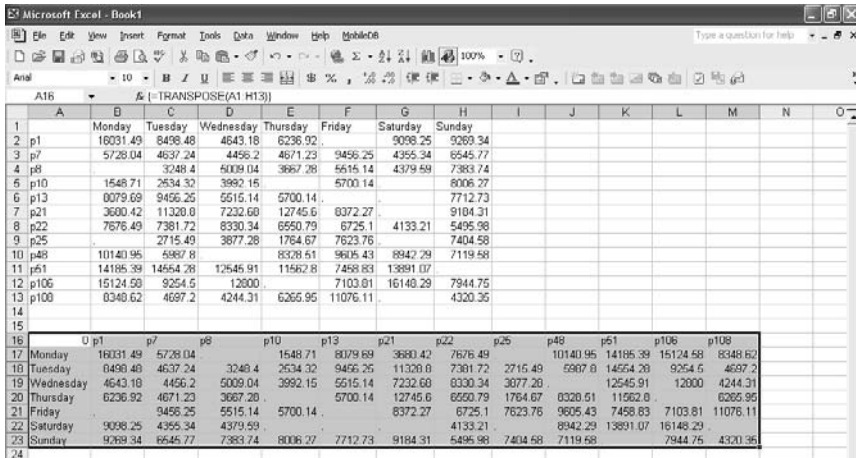
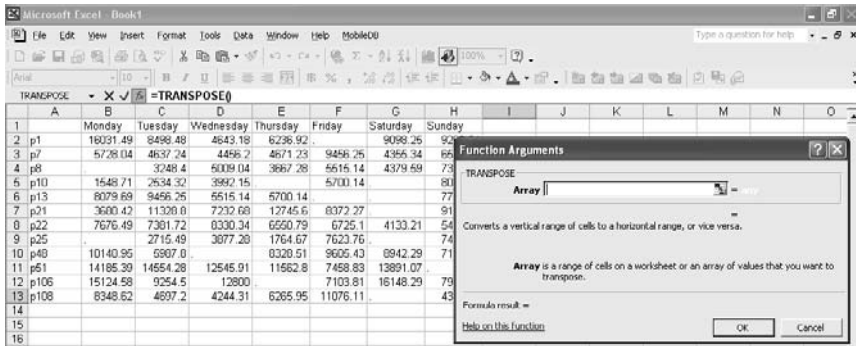
APPENDIX

Steps for Replacing Missing Values

The following is the step-by-step procedure on how to replace missing values. The example used will focus on the Condition C from the II-centered approach.

Step 1. Create a semisimulation data set. The process of creating an artificial data set was explained in the methods section in detail.

Step 2. Transpose the data set. The II-centered approach can be applied by transposing the data set (i.e., matrix) from the variables located in the column to the row (refer to the example below). This task can be easily completed in Microsoft Excel program with a function command “TRANSPOSE.”



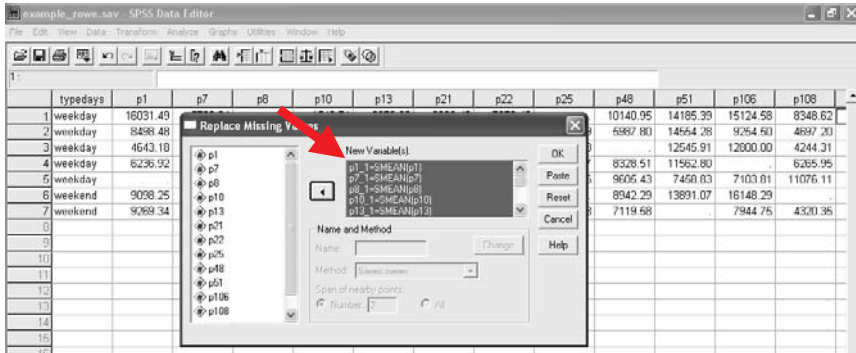
Step 3. Read the data into a SPSS file and recode the variable name. After transposing the data from Excel, read them into a SPSS file and add a new variable (typedays) i.e., code Monday to Friday as “weekday” and Saturday and Sunday as “weekend” (refer to the example below).

typedays	p1	p7	p8	p10	p13	p21	p22	p25	p48	p51	p106	p108
weekday	16031.49	5728.04		1548.71	8079.69	3680.42	7676.49		10140.95	14185.39	15124.58	8348.62
weekday	8498.48	4637.24	3248.40	2634.32	9466.25	11328.80	7381.72	2716.49	6867.80	14654.28	9254.50	4697.20
weekday	4643.18	4466.20	6009.04	3992.16	5515.14	7232.68	8330.34	3697.26		12545.91	12800.00	4244.31
weekday	6236.92	4671.23	3667.28		5700.14	12745.60	6550.79	1764.67	8328.51	11562.80		6265.95
weekday		9496.25	5515.14	5700.14		8372.27	6725.10	7623.76	9605.43	7458.83	7103.81	11076.11
weekend	9098.25	4365.34	4379.59			4133.21			8942.29	13891.07	16148.29	
weekend	9269.34	6545.77	7383.74	8006.27	7712.73	9184.31	5495.98	7404.58	7119.58		7944.75	4320.35

Step 4. Organize output by group. Select the “Split File” function from “Data” menu bar and click “Organize output by group.” This function provides the outputs, organized by “weekday” and “weekend days.”

typedays	p1	p7	p8	p10	p13	p21	p22	p25	p48	p51	p106	p108
1 weekday	16031.49	5728.04		1548.71	8079.69	3680.42	7676.49		10140.95	14185.39	15124.58	8348.62
2 weekday	8498.48								6867.80	14554.28	9254.50	4697.20
3 weekday	4643.18									12545.91	12800.00	4244.31
4 weekday	6236.92								8328.51	11562.80		6265.95
5 weekday									9605.43	7458.83	7103.81	11076.11
6 weekend	9098.25								8942.29	13891.07	16148.29	
7 weekend	9269.34								7119.58		7944.75	4320.35

Step 5. Run “Replace Missing Value.” The missing value replacements can be obtained by “Replace Missing Value” command from SPSS. Choose the “Replace Missing Value” function from “Transform” menu bar, select variables (e.g., p1, p2, ..., p108) from the left window, and move them to new variables.



Step 6. Find the missing value replacements. After clicking “OK” button, you will see the missing value replacements (another set of data) next to your original data set. The new data variable label will have the extension “_1” added to it.

