What Score Should Johnny Get? Missing_Items SAS® Macro forAnalyzing Item Responses On Summative Scales

Patricia Rodriguez de Gil and Jeffrey D. Kromrey

UNIVERSITY OF SOUTH FLORIDA

INTRODUCTION

Should we care about missing data?

Missing data is a pervasive problem in educational research

- Reduce sample size and lose information
- Skew the results and introduce bias into the analysis
- Complicate the interpretation of data analyses
- Reduce statistical power

Particularly, missing data are problematic in research using Likert scales due to item non-response

SIMULATION RESEARCH

A simulation study was conducted to investigate the effectiveness of four imputation procedures:

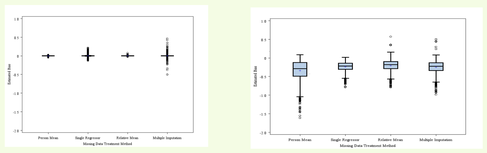
- multiple imputation (***MI***),
- single regression substitution (***SRS***),
- relative mean substitution (***RMS***), and
- person mean substitution (***PMS***),

within summative scales . Table 1 shows the study design factors.

Type	Likert	Shape	Sample Size	Items	Missing P	Missing I	R12
Random N-Rand.	5 7	1 2 3	10 50 100 500	5 10 20 40	.20 .40 .60	.20 .40 .60	.10 .30 .50
2 x	2 x	3 x	4 x	4 x	3 x	3 x	3 =
						Total Conditions	5,184

Table 1. Simulation study on missing data

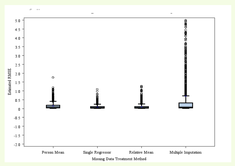
BIAS RESULTS

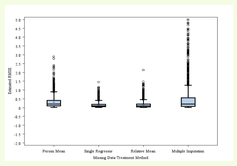


Figures 1 and 2. Distributions of estimates of statistical bias in random and nonrandom missing data conditions

As shown in Figures 1 and 2, for all four methods, nonrandom missing data led to larger, more extreme bias than those for the random condition. MI was particularly a poor estimator when sample sizes were very small.

RMSE RESULTS





Figures 3 and 4. Distributions of RMSE for random and nonrandom missing data conditions

A comparison of the RMSE summaries shows that the four methods generated large sample variability for the nonrandom missing data conditions,

MISSING DATA METHODS

Multiple Imputation (*MI*)

This iterative "fill-in" missing-data method creates a number of m (m>1) imputed data sets, filling each missing value m times using m independent draws, each of which is a plausible estimate of the missing value.

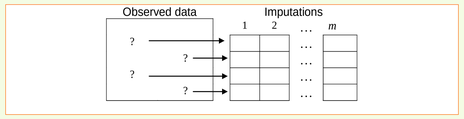


Figure 5. Multiple Imputation, where ***m*** is the number of imputations

Single Regression Substitution (*SRS*)

For a respondent presenting valid responses to items 1 through *a* – 1, but missing data for item *a*, the item that correlates most highly with item *a* is used to predict the missing item response.

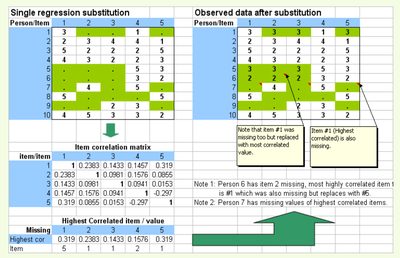


Figure 5. Single regression substitution method

Relative Mean Substitution (RMS)

This method estimates missing data using three sources of information: the person mean of the k^{th} respondent for all valid (nonmissing) item scores, the grand mean of all valid item scores of all respondents, and the mean of all valid scores on the i^{th} item, excluding person *k*. (Raaijmakers, 1999)

$$X_{ak} = \left(\frac{\sum_{i=1}^{n} X_{ik}}{\sum_{j=1}^{N}\sum_{i=1}^{n} X_{ij}} \right) \left(\frac{\sum_{j=1}^{N} X_{aj}}{N} \right); (j \neq k)$$

Person mean substitution (PMS)

This method substitutes the mean of the nonmissing items for person *k* for person *k*'s missing items. Depends on this person's score on all missing items. (PMS only looks at one person and all available items).

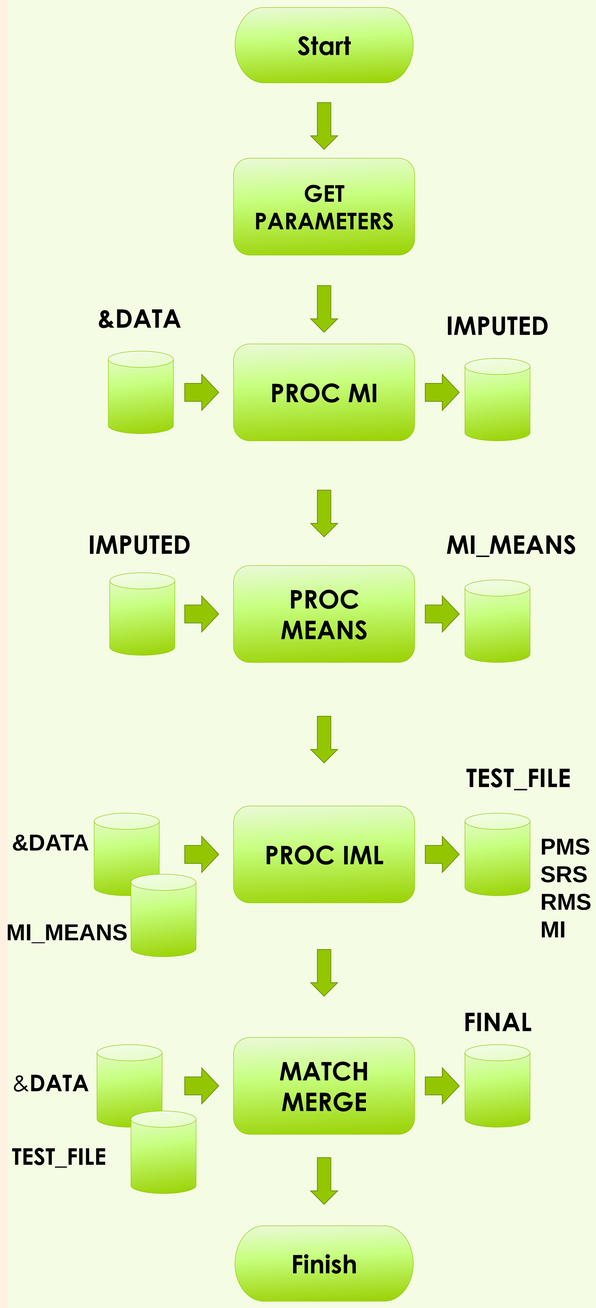
Person/Item	1	2	3	4	5	PMS
1	3	.	.	1	.	(3+1)/2=2
2	2	3	4	4	1	n/a
3	5	2	2	2	5	n/a
4	4	3	2	2	3	n/a
5	.	.	.	5	3	(5+3)/2=4
6	.	.	.	3	2	(3+2)/2=2.5
7	.	4	.	5	.	(4+5)/2=4.5
8	5	.	.	.	5	(5+5)/2=5
9	.	.	2	3	.	(2+3)/2=2.5
10	4	5	3	3	2	n/a

Figure 6. Person mean substitution method

MACRO MISSING_ITEMS

&DATA = _LAST_
&VARS = x1 – x10
&ID = idn
&ROUND=no
&TYPE=mean
&PRINT=yes

FINAL
Original data +
PMS,
SRS,
RMS,
MI



MACRO EXECUTION:AN EXAMPLE

The following SAS data step creates a SAS data set called **ONE**, containing 5 observations. Each observation has responses to 5 individual items (**x1-x5**) as well as an identifier (**ID**). Missing data have occurred on items x4 and x5 for three of the cases.

```
DATA one;
 input ID x1 x2 x3 x4 x5;
cards;
1  1 2 3 4 5
2  2 2 3 . 5
3  3 1 4 2 .
4  3 3 2 . .
5  2 2 1 4 3
;
```

The macro **MISSING_ITEMS** is called twice, to illustrate the difference between computing the total scores as the sum versus the mean.

```
%missing_items (data=one, vars=x1 - x5, id=id,
round = yes, type = mean, print = yes);
%missing_items (data=one, vars=x1 - x5, id=id,
round = yes, type = sum, print = yes);
run;
```

In each call to the macro, the data set is identified (data = one), and both the SAS names for the item variables (vars = X1 – X5) and the case identifier variable (id = id) are provided. In each macro call, rounding is requested (round = yes) and a printed list of cases is requested (print = yes). The only difference in the two macro calls is the method by which total scores will be produced (type = mean in the first macro call, and type = sum in the second).

OUTPUT EXAMPLES

This is data set final

Obs	ID	x1	x2	x3	x4	x5	Person Mean Substitution	Simple Regression Imputation	Relative Mean Substitution	Multiple Imputation	N_Miss
1	1	1	2	3	4	5	3.00	3.0	3.00	3.00	0
2	2	2	2	3	.	5	3.00	3.2	3.00	3.07	1
3	3	3	1	4	2	.	2.50	3.2	2.50	2.85	1
4	4	3	3	2	.	.	2.67	3.6	2.67	3.12	2
5	5	2	2	1	4	3	2.40	2.4	2.40	2.40	0

Output 1. Sample output from first macro call when computing total scores as means

This is data set final

Obs	ID	x1	x2	x3	x4	x5	Person Mean Substitution	Simple Regression Imputation	Relative Mean Substitution	Multiple Imputation	N_Miss
1	1	1	2	3	4	5	15.00	15	15.00	15.00	0
2	2	2	2	3	.	5	15.00	16	15.00	15.33	1
3	3	3	1	4	2	.	12.50	16	12.50	14.33	1
4	4	3	3	2	.	.	13.33	18	13.33	15.67	2
5	5	2	2	1	4	3	12.00	12	12.00	12.00	0

Output 2. Sample output from second macro call when computing total scores as sums

Outputs , as requested, print the estimates of person total scores as sum or means, obtained using multiple imputation (***MI***), single regression substitution (***SRS***), relative mean substitution (***RMS***), and person mean substitution (***PMS***). Total scores in the outputs have been rounded to the nearest hundredth.