**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.

**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](Image)

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](Image)

Relative Mean Substitution (RMS)

This method estimates missing data using three sources of information: the person mean of the jth 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 jth item, excluding person k.

\[ \hat{X}_{ij} = \frac{1}{N} \sum_{k=1}^{N} X_{ik} \]

Person mean substitution (PMS)

This method substitutes the mean of the nonmissing items 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).

![Figure 6. Person mean substitution method](Image)

**MACRO MISSING_ITEMS**

**INPUTS**

- ADATA = _LAST
- &VARS = x1 - x10
- &ID = idn
- &ROUND=no
- &TYPE=mean
- &PRINT=yes

**OUTPUTS**

- 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.

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

```sas
%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);
```

**OUTPUT EXAMPLES**

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

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

**OUTPUT EXAMPLES**

**PERCENTILE**

<table>
<thead>
<tr>
<th>ID</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>Mean Substitution</th>
<th>Relative Mean Substitution</th>
<th>Multiple Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2.00</td>
<td>1.00</td>
<td>2.00</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2.50</td>
<td>2.00</td>
<td>2.50</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>.</td>
<td>2.50</td>
<td>2.00</td>
<td>2.50</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2.40</td>
<td>2.40</td>
<td>2.40</td>
</tr>
</tbody>
</table>

**FINAL**

Outpus, 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.