Imputation for Questionnaires and Scales

Proration (Averaging the Available Items)

Scale scores sum or average a set of questionnaire items measuring the same construct

When component items are missing, researchers often compute prorated scale scores scales by averaging the available item responses

Equivalent to imputing values with a person's mean

Proration = Person-Mean Imputation

Proration

ID	Q1	Q2	Q3	Scale
1	1	2	1	1.3
2	5	NA	4	4.5
3	3	2	4	3.0
4	NA	3	NA	3.0

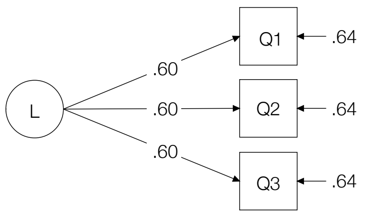
Person-Mean Imputation

ID	Q1	Q2	Q3	Scale
1	1	2	1	1.3
2	5	4.5	4	4.5
3	3	2	4	3.0
4	3.0	3	3.0	3.0

Proration Assumptions

Proration requires two very strict assumptions: identical item means and inter-item correlations (parallel factor structure) and an MCAR mechanism

	Q1	Q2	Q3
Q1	1.00		
Q2	0.36	1.00	
Q3	0.36	0.36	1.00
Means	3.00	3.00	3.00



Scale-Level and Item-Level Imputation

Scale-level imputation: (a) compute scale scores, treating the scale as missing when one or more items is missing, then (b) impute the scale scores directly without the items

Item-level imputation: (a) impute the items, then (b) compute scale scores from the filled-in items

Item-level imputation offers a dramatic power gain

Scale-Level Imputation

Original variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Imputation variables

ID	ScaleX	ScaleY
1	4	NA
2	NA	NA
3	9	7
4	NA	10
…		
200	13	7

Item-Level Imputation

Original variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Imputation variables

ID	X_1	X_2	X_3	Y_1	Y_2
1	1	2	1	NA	3
2	5	NA	4	NA	NA
3	3	2	4	3	4
4	NA	3	NA	5	5
…					
200	4	5	4	3	4

Motivating Example

Questionnaire data from a study of eating disorder risk in a sample of 500 college-aged women

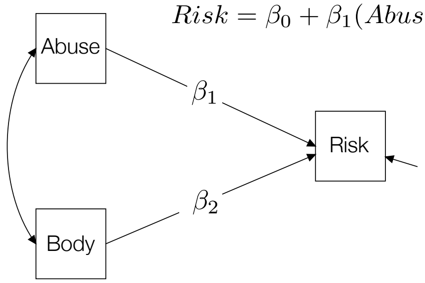
Variables include body mass index (BMI), questionnaire items measuring body dissatisfaction and eating disorder risk, past sexual abuse history (0 = no abuse history, 1 = abuse history)

All questionnaire items measured on a 7-point scale

Analysis Model

Body dissatisfaction and eating disorder risk are scale scores computed as the sum of the item responses

$$Risk = \beta_0 + \beta_1(Abuse) + \beta_2(Bodydis) + e$$



Analysis Summary

Item-Level Imputation

	Est.	SE	z	p	RIV
B_0	10.04	0.88	11.36	< .001	0.04
B_1 (Abuse)	1.36	0.51	2.69	0.01	0.25
B_2 (Diss)	0.50	0.03	16.27	< .001	0.04

Scale-Level Imputation

	Est.	SE	z	p	RIV
B_0	9.45	1.14	8.27	< .001	0.80
B_1 (Abuse)	1.70	0.60	2.81	< .01	0.98
B_2 (Diss)	0.52	0.04	12.66	< .001	0.86

Important Takeaways

Item-level imputation offers a dramatic gain in precision

RIV = .86 (relative increase in variance) means that missing data error was 86% as large as complete-data sampling error, versus only 4% for item-level imputation

In order to obtain the same standard errors as item-level imputation, scale-level imputation would require a 40% to 90% increase in the sample size

Ex6.1.imp
Blimp Diagnostic Script

```
DATA: ~/desktop/examples/eatingrisk.csv;
VARNAMES: id abuse bmi bds1-bds7 edr1-edr6;
NOMINAL: abuse;
ORDINAL: bds1-bds7 edr1-edr6;
MISSING: -99;
MODEL: ~ abuse bmi bds1-bds7 edr1-edr6;
SEED: 90291;
BURN: 20000;
THIN: 1;
NIMPS: 2;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```

Diagnostic Output

```
Comparing iterations 9251 to 18500 for 2 chains.
                      ----------------------------------------------------
                      |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                      ----------------------------------------------------
             Max PSR  |       1.007|         nan|       1.000|       1.055|
    Missing Variable  |        edr2|            |         bmi|        bds1|
                      ----------------------------------------------------

Comparing iterations 9301 to 18600 for 2 chains.
                      ----------------------------------------------------
                      |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                      ----------------------------------------------------
             Max PSR  |       1.007|         nan|       1.000|       1.046|
    Missing Variable  |        edr2|            |         bmi|        bds1|
                      ----------------------------------------------------
```

Ex6.2.imp
Blimp Imputation Script (Mplus Format)

```
DATA: ~/desktop/examples/eatingrisk.csv;
VARNAMES: id abuse bmi bds1-bds7 edr1-edr6;
NOMINAL: abuse;
ORDINAL: bds1-bds7 edr1-edr6;
MISSING: -99;
MODEL: ~ abuse bmi bds1-bds7 edr1-edr6;
SEED: 90291;
BURN: 10000;
THIN: 10000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Blimp Output

```
---------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

   id abuse bmi bds1 bds2 bds3 bds4 bds5 bds6 bds7 edr1 edr2
   edr3 edr4 edr5 edr6

---------------------------------------------------------------
```

Ex6.3.inp
Mplus Analysis Script

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = id abuse bmi bds1-bds7 edr1-edr6;
usevariables = abuse bodydis eatrisk;
DEFINE:
bodydis = sum(bds1-bds7);
eatrisk = sum(edr1-edr6);
MODEL:
eatrisk on abuse bodydis;
OUTPUT:
standardized(stdyx);
```

Mplus Analysis Output

```
MODEL RESULTS

                                                    Two-Tailed   Rate of
                    Estimate       S.E.  Est./S.E.    P-Value    Missing

 EATRISK  ON
    ABUSE              1.355      0.512      2.646      0.008      0.206
    BODYDIS            0.500      0.031     16.188      0.000      0.047

 Intercepts
    EATRISK           10.092      0.879     11.479      0.000      0.032

 Residual Variances
    EATRISK           12.142      0.800     15.170      0.000      0.080
```

Ex6.4.imp
Blimp Imputation Script (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/eatingrisk.csv;
VARNAMES: id abuse bmi bds1-bds7 edr1-edr6;
NOMINAL: abuse;
ORDINAL: bds1-bds7 edr1-edr6;
MISSING: -99;
MODEL: ~ abuse bmi bds1-bds7 edr1-edr6;
SEED: 90291;
BURN: 10000;
THIN: 10000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Blimp Output

```
-------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

  imp# id abuse bmi bds1 bds2 bds3 bds4 bds5 bds6 bds7 edr1
  edr2 edr3 edr4 edr5 edr6

-------------------------------------------------------------
```

Ex6.5.r
R Analysis Script

```
# Required packages
library(mitml)

# Read data
impdata <- read.csv(file = ”~/desktop/examples/
  imps.csv”,head = FALSE, sep = ”,”)
names(impdata) = c(”imp”,”id”,”abuse”,”bmi”,
  ”bds1”,”bds2”,”bds3”,”bds4”,”bds5”,”bds6”,”bds7”,
  ”edr1”,”edr2”,”edr3”,”edr4”,”edr5”,”edr6”)

# Compute scale scores
impdata$bodydis <- rowSums(subset(impdata,
  select = bds1:bds7))
impdata$eatrisk <- rowSums(subset(impdata,
  select = edr1:edr6))
```

Ex6.5.r
R Analysis Script

```
# Analyze data and pool estimates
implist <- split(impdata, impdata$imp)
implist <- as.mitml.list(implist)
analysis <- with(implist, lm(eatrisk ~ abuse + bodydis))
testEstimates(analysis, df.com = 497)

# Test full model with Wald test
emptymodel <- with(implist, lm(eatrisk ~ 1))
testModels(analysis, emptymodel, method = ”D1”)
```

R Analysis Output

```
Final parameter estimates and inferences obtained from 20 imputed data sets.

             Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
(Intercept)    10.092     0.882    11.445   467.989     0.000     0.032     0.031
abuse           1.355     0.513     2.640   213.738     0.009     0.253     0.205
bodydis         0.500     0.031    16.141   447.769     0.000     0.049     0.047


Model comparison calculated from 20 imputed data sets.
Combination method: D1

    F.value      df1      df2    P(>F)      RIV
    155.593        2 2010.888    0.000    0.142

Unadjusted hypothesis test as appropriate in larger samples.
```

Ex6.6.sps
SPSS Analysis Script

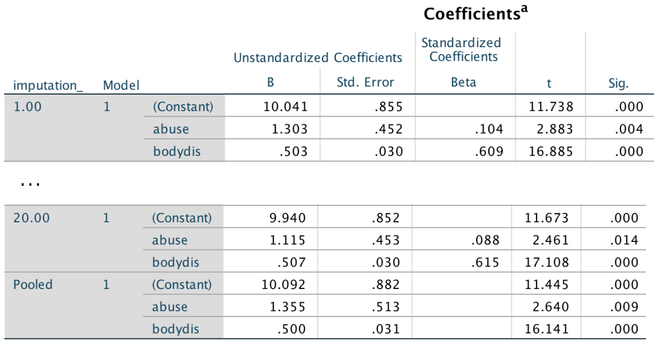
```
data list free file = ‘/users/craig/desktop/examples/imps.csv’
 /imputation_ id abuse bmi bds1 to bds7 edr1 to edr6.
exe.

* Compute scale scores
compute bodydis = sum(bds1 to bds7).
compute eatrisk = sum(edr1 to edr6).
exe.

* Initiate pooling routines.
sort cases by imputation_.
split file layered by imputation_.

* Analysis and pooling.
regression
  /descriptives mean stddev corr sig n
  /dependent eatrisk
  /method=enter abuse bodydis.
```

SPSS Analysis Output



Coefficients[a]

			Unstandardized Coefficients		Standardized Coefficients		
imputation_	Model		B	Std. Error	Beta	t	Sig.
1.00	1	(Constant)	10.041	.855		11.738	.000
		abuse	1.303	.452	.104	2.883	.004
		bodydis	.503	.030	.609	16.885	.000
...							
20.00	1	(Constant)	9.940	.852		11.673	.000
		abuse	1.115	.453	.088	2.461	.014
		bodydis	.507	.030	.615	17.108	.000
Pooled	1	(Constant)	10.092	.882		11.445	.000
		abuse	1.355	.513		2.640	.009
		bodydis	.500	.031		16.141	.000

Ex6.7.do
Stata Analysis Script

```
// Import and save original data
import delimited ”~/desktop/examples/eatingrisk.csv”
rename (v1 - v16)(id abuse bmi bds1 bds2 bds3 bds4 bds5 bds6 bds7 edr1
  edr2 edr3 edr4 edr5 edr6)
generate imp = 0

// Recode missing values
foreach var of varlist id - edr6 {
      replace ‘var’ = . if ‘var’== -99
}
save original, replace

// Import and save imputed data
clear
import delimited ”~/desktop/examples/imps.csv”
rename (v1 - v17)(imp id abuse bmi bds1 bds2 bds3 bds4 bds5 bds6 bds7
  edr1 edr2 edr3 edr4 edr5 edr6)
save imputed, replace
```

Ex6.7.do
Stata Analysis Script

```
// Append original and imputed data
use original, clear
append using imputed

// Compute scale scores
generate bodydis = bds1 + bds2 + bds3 + bds4 + bds5 + bds6 + bds7
generate eatrisk = edr1 + edr2 + edr3 + edr4 + edr5 + edr6

// Convert to mi data
mi import flong, m(imp) id(id) imputed(abuse bmi bds1 - bds7
  edr1 - edr6 bodydis eatrisk) clear

// Analyze data and pool results
mi estimate, cmdok: regress eatrisk abuse bodydis
```

Stata Analysis Output

```
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =        500
                                                Average RVI       =     0.1078
                                                Largest FMI       =     0.2062
                                                Complete DF       =        497
DF adjustment:   Small sample                   DF:     min       =     213.74
                                                        avg       =     376.50
                                                        max       =     467.99
Model F test:       Equal FMI                   F(   2,  395.6)   =     155.59
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
     eatrisk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       abuse |   1.355301     .513388     2.64   0.009     .3433494    2.367253
     bodydis |   .5000707    .0309808    16.14   0.000     .4391848    .5609566
       _cons |   10.09217     .881764    11.45   0.000     8.359459    11.82487
------------------------------------------------------------------------------
```