Disparities Among Children Served by the CMHS Children's Services Program

Overview of Multiple Imputation and Using Multiply Imputed Data

Melissa Azur, Constantine Frangakis, and Elizabeth Stuart

Updated May 2, 2008

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Why do we need to use multiple imputation?
In nearly all surveys there will be some missing values, and it is a particular concern in longitudinal studies. In the National Evaluation of the CMHS Children and Their Families Program, 12-month follow-up rates for Phase II and Phase III sites varied from 93% to 0%. In the CMHS disparities project, this includes individuals who did not respond to all items as well as entire sites that did not administer all of the instruments. An analysis that simply drops all individuals with incomplete data can lead to bias if there are differences in the individuals with and without missing values (i.e., data not missing completely at random), and can be inefficient because it does not utilize all available data. A simple approach to dealing with the missingness is to fill in (or "impute") the missing values. However, analyses that use standard single imputation (imputing one value for each missing value) will underestimate the true variability in the data because they do not account for the fact that the imputed values are not in fact the true, observed data. Single imputation will thus result in variance estimates that are biased towards zero, which will lead to anti-conservative results in terms of higher significance levels (lower p-values, shorter confidence intervals) than are valid.

Multiple imputation addresses this problem by generating multiple imputations for each missing value (Rubin 1987; Schafer 1997). For example, if a participant's internalizing score is missing, we generate 5 predicted values for that score, using other information observed on the participant (e.g., age, race, sex, etc.). This creates 5 "complete" datasets, where within each dataset all missing values are filled in.

How are the imputations generated?
To generate the imputed values we use a procedure called "Multiple Imputation by Chained Equations" (MICE; Raghunathan et al. 2001). The missing values for a given participant are predicted from that participant's observed values, using relations observed in the data for other participants. This is accomplished by running a series of regression equations. In each regression, a variable with missing values is regressed upon other variables used in the imputation process. The resulting regression model is used to predict the missing values for that variable. The process then moves to the next variable with missing data, where that variable is regressed upon all the other variables and previously generated predictions. The process continues in an iterative manner until all missing values have been imputed.

A benefit of using the MICE procedure is that the regression models reflect the types of variables being imputed, with, for example, binary variables modeled using logistic regression and continuous variables modeled using normal linear regression.

Details of Imputations for CMHS Disparities Project
Data from the National Evaluation Outcome Study for Phases II & III were imputed using the MICE procedures described above. IVEware (University of Michigan, 2002), free software that runs through SAS, was used to generate the imputations. Given the size of the dataset and space/memory limitations, certain decisions were made regarding which variables to impute and how to specify the imputation model. For example, for variables with both raw and standard scores available, the raw scores were imputed rather than the standard scores (or both). Analysts who choose to impute raw scores, but plan to analyze standard scores, will want to make sure they have the capability to transform the raw scores to standard scores after the imputations have been completed. Appendix A lists the variables included in the imputation model. Rather than including all the variables in the dataset as predictors in the regression models, stepwise regression with

forward selection was used to select the predictors in each model. The predictors available for inclusion in each model were 1) all other variables in the dataset, 2) site indicators, and 3) selected two-way interactions between variables in the data. For each variable, the stepwise selection model picked as predictors those variables that were the most predictive of the variable being imputed. For example, when imputing the variable "sex problems," the stepwise-selection model chose the variables "child has been sexually abusive to others," "sexual acting out," "CBCL total behavior problem score," and "CBCL delinquent behavior score" as most predictive of sex problems. Of the 400 variables possible for inclusion as predictors in the regression models, an average of 6 variables was selected in each regression.

In ideal circumstances, the imputation model should be general enough to preserve relationships between variables that are of interest in the analysis stage. This means that all variables that might be used in the analytical model should first be included in the imputation model. Variables that might *not* be included in the analytical model, but that have an association with variables in the analytical model, should also be included in the imputation model. This concept applies to interactions, such that under ideal conditions, interactions that are included in the analytical model should first be included in the imputation model. An analytic model that includes interactions that were not present in the imputation model may result in estimates that are biased toward the null. In other words, differences may not be detected. In the imputed dataset we created, the imputation model includes all variables listed in Appendix A. Due to constraints in the size of our model, it was not possible to include all potential interactions in the dataset. Interactions between our core disparities variables (age, gender, race) were included, as well as interactions between these variables and poverty and referral source. Using guidelines from the multiple imputation literature, 5 imputed datasets were created.

Assumptions

Creation of the imputed data relies on a series of assumptions. First, the missing data are assumed to be Missing at Random (MAR). This is a weaker assumption (and more realistic) than Missing Completely at Random (MCAR). MCAR says that the probability of a value being missing is the same for all individuals. In other words, the children with a missing value on a given variable are a random sample from all children. This implies that if a child has a missing value for "CBCL total behavior problem score," then we could impute that value from any other child's observed value. MCAR is what is assumed by simple missing data techniques that just use cases with observed data (and do no imputations) or that do a simple mean imputation. It is a generally fairly unrealistic assumption. We instead assume MAR, which says that the missing data mechanism is not completely random, but depends only on values that are observed—not on the missing values themselves. In other words, that the children with a missing value on a given variable are a random sample from all children with the same values on the observed variables. This implies, for example, that to impute a child's missing "CBCL total behavior problem score" we could use data from other children with the same observed characteristics as the child with the missing data. Most standard multiple imputation techniques rely on an assumption of MAR.[1]

[1] If the missing data mechanism depends on the unobserved values themselves, it is said to be Not Missing at Random (NMAR), which requires more complex modeling approaches. This would occur if, for example, even among children with the same values on the observed data, children with high behavior problem scores were less likely to have their problem score reported.

Second, we do the imputations with all sites pooled together. This relies on an assumption that, generally, it is better to impute missing values using children with similar observed characteristics from other sites than it is to impute missing values using children in the same site, who may have very different observed characteristics than the child with a missing value. This also assumes that, for example, if we need to impute a child's "CBCL total behavior problem score," the models will essentially use other children with the same age, race, gender, etc. as the child with the missing value, even if they might be from other sites. The exception to this is that if being in a particular site is very predictive of a variable (e.g., children in the West Chester County site have more previous psychiatric hospitalizations than children in other sites), then the stepwise procedure will likely select that site as one of the predictors. This will allow the imputation of that variable to depend on whether a child is in that site or not. In this way, the data (and models) determine which sites should not be pooled by allowing site indicators to appear in the prediction models as appropriate.

Look at all the data I can use!

Team members may want to jump in and use the imputed data. As with any project, it is important to spend some time understanding the data before beginning analyses. As mentioned earlier, almost all of the variables in the dataset had at least some values imputed. This includes variables that had little missingness and variables that had complete missingness. We have created a document to be used in conjunction with this manual that summarizes information regarding frequency of missingness and specific patterns of imputed data. For example, information is provided on the amount of missingness by variable in the observed data and variables that were completely missing for particular sites in the observed data. The document also provides a summary of variables where there may be concerns regarding how well the imputation model imputed the data. That document should be used to guide decision making regarding which variables to include in analyses and whether some sites should be excluded from analyses.

Why is there still missing data in the imputed dataset?

Despite the fact that we imputed the data, you will notice that there are still missing values in the dataset. This is due to two reasons. First, some variables, such as DSM axis 1 diagnostic code, were carried over to the imputed dataset but were not used in the imputation model. Please refer to Appendix A for a list of variables included in the imputation model. Secondly, there is missingness in the imputed data by design. There are instances where an imputed value is inappropriate and we accounted for this in the specification of the imputation model. For example, if a youth reported never using cocaine, he/she should not have a value for *age of first use*.

Where to start?

Some people find it helpful to begin data analyses by working with one imputed dataset. The Stata commands are easily applied to the multiply imputed data (across the five imputed datasets), but depending upon the type of model you are running and the length of time involved in running the model, it may be more efficient to initially work with one imputed dataset to develop your analytical models. Once you have a sense of the type of models you want to run, you can apply them to the multiply imputed data. When running preliminary models using just one dataset it is important to not rely very heavily on the standard errors and significance levels obtained, as the standard errors will almost always be too small. Any final inferences should be made using all 5 imputed datasets, as described below.

How do I analyze multiply imputed data?

Once the data have been imputed, each imputed dataset is “complete” in the sense that it has no missing values (except those missing by design, as discussed above). Analyzing multiply imputed data involves two steps: 1) running standard analyses (e.g., regression) on each of the imputed datasets, and 2) combining the estimates from each dataset to obtain the final result. The variance estimates in Step 2 involve calculating both the “within” variance calculated for each dataset individually, as well as the “between” variance that reflects the uncertainty in the imputations—how variable the results are across the 5 imputed datasets. Schafer (1997) describes the details of how results are combined across the imputed datasets.

While it is possible to do this combining yourself (e.g., by writing a short computer program), many standard statistical software packages now include procedures to combine results across datasets automatically.[2] Appendix C provides information on how to analyze multiply imputed data in SAS, R, SPSS, HLM, and Mplus. The procedures for analyzing imputed data in Stata are described below.

Analyzing Multiply Imputed Data in Stata

Stata has three sets of imputation commands available for use when analyzing multiply imputed data. These commands are available as ado files and must be downloaded from Stata. (In the command window, type *findit* name of command). It is strongly recommended that users review the help file for each of these commands prior to use.

The first suite of commands includes *mimstack* and *mim*. The command *mimstack* prepares the data for the *mim* command. It stacks the imputed datasets on top of each other and creates two new variables, one that identifies each unique observation and one that identifies each imputed dataset. The default option of this command stacks the unimputed and imputed datasets together. While this is a useful option for comparing imputed and non-imputed data, for our purposes this is unnecessary. An example of the *mimstack* command is illustrated below.

```
mimstack, m(5) so("childid") nomj0 istub(impset)
```

As you can see, the command includes a number of different components. The *m( )* indicates the number of imputed datasets. The *so* specifies the variable that uniquely identifies the observations. In our case this is the variable “childid.” The *nomj0* tells Stata that the original (non-imputed) dataset is not to be stacked with the imputed datasets. The *istub* specifies the name of the datasets to be stacked. Please refer to the *mimstack* Stata help file for more information. Given the size of the datasets, you will either need to increase the memory in Stata prior to running *mimstack* or create smaller datasets that includes only the variables of interest for your analyses. We recommend creating smaller datasets.

Once you have prepared the datasets using *mimstack*, you are ready to conduct analyses using the *mim* command. This command is used with estimation commands (e.g., regress) and can also be used for data manipulation and post-estimation analyses.

[2] Please refer to Appendix B for the formula to combine results across imputed datasets.

The *mim* command provides point estimates, overall variance estimates, and confidence intervals. The *mim* command can be used with a wide range of estimation commands, including *regress*, *logit*, *ologit*, *proportion*, *mean*, *xtlogit*, *xtreg*, *xtpoisson*, and *glm* (see the *mim* Stata help file for a full list of estimation commands). The *mim* command also has an option that will output the estimation results for each individual dataset. In addition to using *mim* with estimation commands, it can be used with the post-estimation commands *lincom*, *testparm*, and *predict*.

An example of *mim* and sample output is included below.

```
    mim: regress anxdepr sex age
```

```
Multiple-imputation estimates (regress)                  Imputations =       5
Linear regression                                        Minimum obs =    8897
                                                         Minimum dof =    31.0

------------------------------------------------------------------------------
     anxdepr |     Coef.  Std. Err.      t    P>|t|    [95% Conf. Int.]   MI.df
-------------+----------------------------------------------------------------
         sex |   .898506   .172013     5.22   0.000     .54767  1.24934    31.0
         age |  -.062865   .019997    -3.14   0.002   -.102137 -.023594   602.1
       _cons |   10.5501   .257587    40.96   0.000    10.0434  11.0568   332.2
------------------------------------------------------------------------------
```

The second suite of commands includes *miset, mifit, milincom*, and *mireset*. The command *miset* creates temporary copies of the imputed datasets so that subsequent commands can be run on the data. This command must be run prior to running other commands in the *miset* suite. The command *mifit* is used with an estimation command (e.g., *regress*). The *mifit* command provides multiple-imputation point estimates, overall variance estimates, and confidence intervals. The *mifit* command can be used with *regress, logit, probit, poisson, glm*, or *xtgee* (see the Stata *mifit* help file for a full list of available estimation commands). The *mifit* command also has an option that will output the estimation results for each imputed dataset separately (by default it outputs only the summary across all 5 imputed datasets). The command *milincom* tests linear combinations of coefficients. The command functions in the same way as the standard *lincom* command. The command *mirest* erases the temporary files that were created by *miset*. An example of *miset* and *mifit* and sample output is provided below.

```
      miset using impset[3]
      mifit: regress anxdepr sex age
```

```
Overall estimates

                                                  Number of obs   =       8897
------------------------------------------------------------------------------
     anxdepr |      Coef.  Std. Err.     t   P>|t|  [95% Conf. Interval]  MI.df
-------------+----------------------------------------------------------------
         sex |   .89851    .17201     5.22  0.000  .54775    1.2493      31.14
         age |  -.06287       .02    -3.14  0.002 -.10213    -.0236     649.80
```

[3] In the above code, if your directory is not set to the datasets' location, you will need to specify the entire path name, e.g. "C:\myoutdir\impullni.dta", rather than simply "impfullni".

```
       _cons |    10.55    .25759    40.96  0.000  10.043    11.057     346.73
------------------------------------------------------------------------------
```

The third suite of Stata commands mirrors the first set of commands and includes *mijoin*, *micombine*, and *misplit*. The m*ijoin* command combines the datasets by stacking them vertically. This command must be used prior to using any of the other commands in the *mijoin* suite. The command *micombine* is used with an estimation command and is similar to *mifit*. It runs analyses on each dataset separately and then combines results. The command provides multiple-imputation point estimates, overall variance estimates, and confidence intervals; it can be used with *regress, logit, probit, poisson,* and *xtgee* (see Stata command help file for a full list of estimation commands). It too has an option to obtain the estimates of each imputed dataset. An example of *mijoin* and *micombine* is provided below.

```
      mijoin impset, clear m(5)

micombine regress anxdepr sex age
```

```
Multiple imputation parameter estimates (5 imputations)
------------------------------------------------------------------------------
     anxdepr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .8985059   .1720133     5.22   0.000     .5613202    1.235692
         age |  -.0628655   .0199967    -3.14   0.002    -.1020636   -.0236673
       _cons |   10.55009   .2575868    40.96   0.000     10.04516    11.05502
------------------------------------------------------------------------------
9185 observations.
```

You will notice that the *mimstack*/*mim* commands, *miset/mifit* commands, and the *mijoin/micombine* commands produce the same output, and in general any of the sets of commands will work. However, in using these commands, we have noticed two differences. When collinearity exists, *mifit* will not run. Users can address this problem by removing collinear variables from the model. The *mim* command and the *micombine* command automatically drop the collinear variables and will run the model. Secondly, *mim* appears to have the greatest capabilities in terms of estimation and post-estimation commands. For example, *mim* is the only command that permits the use of random effects models. Thus, in analyses that take into account clustering by site, the *mim* command is the only available option at this time. Users who choose to utilize the *mifit* and *micombine* commands can switch back and forth between the two commands by using *mijoin* and *misplit*. Please refer to the Stata help files for more information.

Conclusion

This document provides an overview of why and how to use multiple imputation, and in particular, the implications for the CMHI disparities analyses. Using multiply imputed data will allow us to take full advantage of the data that we have, while also accounting for the uncertainty that does exist because of the missing data. We hope that this document is helpful in carrying out your analyses. Please see the references below or contact us if you have any additional questions.

Where can I get more information?

www.multiple-imputation.com
http://oregonstate.edu/~acock/growth-curves/royston_SJ_paper_nearly_final.pdf
http://www.stat.psu.edu/~jls/mifaq.html

Little, R. J. A. and D. B. Rubin (1987). Statistical analysis with missing data. John Wiley & Sons, New York.

Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of state of the art. *Psychological Methods* 7(2): 147-177.

References

Barnard, J. and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* 86(4): 948-955.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27, 85-95.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data*. Chapman & Hall, London. Book No. 72, Chapman & Hall series Monographs on Statistics and Applied Probability.

University of Michigan. (2002). IVEware: Imputation and Variance Estimation Software. University of Michigan. http://www.isr.umich.edu/src/smp/ive/

Appendix A: Variables Used in the Imputation Model

siteid
nrace1
nrace2
nrace3
nrace4
nrace5
nrace6
nrace7
nrace8
age
susa1
susa2
susa3
susa4
susa5
susa5a
susa14d
susa15d
susa18d
susa21d
totadu
totchild
relation
custody
ref1
diqintv
diqlang
sex
ref2
curserv
srvoutp
srvschol
srvdytr
srvrtcih
srvalch
psych
pabuse
sabuse
run
caregiver
suicide
drug
sexabuse
famvil
abuse
famill
family
psycychp
pfealony
famabu
parntab
respsex
livhom
livmon
livdays
income
respage
Medicaid
paysrvs
chronic
medphys
medbeh
prob1
prob2
prob3
prob4
prob5
prob6
prob7
prob8
prob9
prob10
prob11
prob12
prob13
prob14
prob15
prob16
prob17
prob18
prob19
prob20
prob21
prob22
prob23
prob24
prob25
prob26
prob27
prob28
prob29
prob30
prob31
prob32
prob33
caflang
brresp
brlang
berist
berfit
beriast
bersft
berast
bersraw
cbcresp
cbcintv
cbcmeth
cbclang
csex
cage
cactivr
csociar
cscholr
ctotcomr
withdrr
somcomr
anxdepr
socpror
thopror
attpror
delbehr
aggbehr
sexpror
ctotraw
cintraw
cextraw
susintv
susmeth
suslang
susa6
susa7
susa8
susa9
susa10a
susa10b
susa10c
susa10d
susa11a
susa11b
susa11c
susa11d
susa12a
susa12b
susa12c
susa12d
susa13a
susa13b
susa13c
susa13d
susa14a
susa14b
susa14c
susa15a
susa15b
susa15c
susa16a
susa16b
susa16c
susa16d
susa17a
susa17b
susa17c
susa17d
susa18a
susa18b
susa18c
susa19a
susa19b
susa19c
susa19d
susa20a
susa20b
susa20c
susa20d
susa21a
susa21b
susa21c
susb1a
susb1b
susb2a
susb2b
susb3a
susb3b
susb4a
susb4b
susb5a
susb5b
susb6a
susb6b
susb7a
susb7b
susb8a
susb8b
susb9a
susb9b
susb10a
susb10b
susb11a
susb11b
susb12a
susb12b
susb13a
susb13b
susb14a
susb14b
susb15a
susb15b
susb16a
susb16b
susb17a
susb17b
susb18a
susb18b
susb19a
susb19b
susb20a
susb20b
susb21a
susb21b
susnum
alc6
cig6
mar6
zipclass
entryear
funding
yrfunded
diqeducg
srolecat
homecat
comcat
behcat
moodcat
harmcat
subcat
thinkcat
diqresp
diqmeth
brintv
brmeth
rolresp
rolintv
rolmeth
rollang
liv1a
totliv
lvre1
eqresp
eqintv
eqmeth
eqlang
edu1
edu1a
edu2
edu3
edu3b
edu3c
edu4
edu4a
edu5
edu5a
edu5b
edu5c
edu6a
edu7
edu7a
edu7c
edu8
edu9
edu10
edu11
edu12

```
edu13
edu14
edu15
edu15a
edu15b
edu16
edu16a
edu17
edu17a
edu17b
edu18
edu19a
edu20
edu21
edu61
edu62
edu63
edu64
edu65
edu66
edu67
edu68
edu69
edu610
edu611
edu7b1
edu7b2
edu7b3
edu7b4
edu7b5
edu7b6
edu7b7
speced
ysrintv
ysrmeth
ysrlang
ysex
yage
yactivr
ysociar
schooly
ytotcomr
yschool
withdry
somcomy
anxdepy
socproy
thoproy
attproy
delbehy
aggbehy
desidry
ytotraw
yintraw
yextraw
ysrresp
resp2rel
resp2sex
concern1
concern2
dsmsub
dsmvco
dsmoth
dsmpsy
dsmmood
dsmaut
dsmanx
dsmadj
dsmpts
dsmimp
dsmodd
dsmadd
dsmcon
dsmdbd
dsmper
dsmmr
dsmld
axis41
axis42
axis43
axis44
axis45
axis46
axis47
axis48
axis49
dsyear
dsintv
dsmeth
dslang
ds1
ds2
ds3
ds4
ds4a
ds5
ds6
ds7
ds7a
ds8
ds9
ds10
ds11
ds12
ds13
ds14
ds15
ds16
ds17
ds18
ds19
ds20
ds20a
ds20b
ds21
ds21a
ds21b
ds21c
ds22
ds22a
ds22b
ds23
ds23a
ds23b
ds23c
ds24
ds24a
ds25
ds25a
```

Appendix B: Formulas for Combining Results Across Datasets

These are sometimes called "Rubin's combining rules." The information below is from Joe Schafer's multiple imputation FAQ list (http://www.stat.psu.edu/~jls/mifaq.html).

Rubin (1987) presented this method for combining results from a data analysis performed *m* times, once for each of *m* imputed data sets, to obtain a single set of results. From each analysis, one must first calculate and save the estimates and standard errors. Suppose that $\hat{Q}_j$ is an estimate of a scalar quantity of interest (e.g. a regression coefficient) obtained from data set *j* (*j*=1,2,...,*m*) and U_j is the standard error associated with $\hat{Q}_j$. The overall estimate is the average of the individual estimates,

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j.$$

For the overall standard error, one must first calculate the within-imputation variance,

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j\,.$$

and the between-imputation variance,

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \bar{Q}\right)^2.$$

The total variance is

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B.$$

The overall standard error is the square root of *T*. Confidence intervals are obtained by taking the overall estimate plus or minus a number of standard errors, where that number is a quantile of Student's t-distribution with degrees of freedom

$$df = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2.$$

A significance test of the null hypothesis *Q*=0 is performed by comparing the ratio $t = \bar{Q}/\sqrt{T}$ to the same t-distribution. These basic combining rules assume a large sample size in the complete datasets. Some procedures (e.g., Stata's *mifit*) instead use formulas developed by Barnard and Rubin (1999) which are more appropriate for small sample sizes. Additional methods for combining the results from multiply imputed data are reviewed by Schafer (1997, Ch. 4).

Appendix C: Details of Statistical Software for Analyzing Multiply Imputed Data

1. SAS: PROC MIANALYZE

http://www.sas.com/rnd/app/papers/mianalyzev802.pdf

This procedure combines the results from the multiply imputed datasets to generate overall results.

Data: This procedure requires that the data is in a form such that the 5 imputed datasets are "stacked" to create one dataset with 5*N observations, where N is the number of participants in the original data. This dataset should contain a variable called "Imputation" which takes values from 1-5 and says which imputation number each observation is from.

Two analysis steps are then required:
Step 1: Run separate analyses (e.g., regression) on each dataset and save the resulting parameter estimates. For example:

```
proc reg data=outmi outest=outreg covout noprint;
          model Oxygen= RunTime RunPulse;
          by _Imputation_;
run;
```

Print out the parameter estimates:

```
proc print data=outreg(obs=8);
          var _Imputation_ _Type_ _Name_ Intercept RunTime RunPulse;
          title 'Parameter Estimates from Imputed Data Sets';
run;
```

Step 2: Use "PROC MIANALYZE" to combine the results:

```
proc mianalyze data=outreg;
      var Intercept RunTime RunPulse;
run;
```

Basically any analysis procedure can be used in Step 1 and the combining rules used in Step 2 will be the same.

2. R: mitools package

cran.r-project.org/doc/packages/mitools.pdf

Data: The data needs to be in an "imputationList" object. See the documentation for more details on how to create this. Essentially, it is a list object, with each dataset as an element.

Again, two analysis steps are then required.
Step 1: Run models on each dataset
This is done using the "with" command. For example, if "implist" is an ImputationList object, use the command with(implist, function) where "function" is a function taking a dataframe argument, such as regression. For example, from the documentation: "models <- with(smi, glm(drinkreg ~ wave*sex, family=binomial())" will run a logistic regression on the "smi" imputationList.

Step 2: Combine results across datasets
The “micombine” command will combine results across the results from the “with” command. For example, “MIcombine(models)” will combine the results from the 5 analyses in Step 1.

See the documentation of the mitools package for more information.

3. SPSS

The following information was obtained from the SPSS technicians.

1) **SPSS External Programmability**: Download the Python module rubin.py from http://www.spss.com/devcentral/index.cfm?pg=downloadDet&dId=55 to calculate statistics from the multiply imputed data and combine them using Rubin's Rules.
2) **More Detail**: See Examples 30 & 31 in the **AMOS 7 User’s Guide**. They show how the multiple imputed data sets can be created in AMOS. The second step, the analysis of each imputed file, could be performed in SPSS or AMOS, depending on the type of analysis. If the analysis was available in SPSS and the imputed data sets had been stacked into one file (which AMOS can do) then a split file structure could be used to analyze each of the imputed data sets with one run, saving the desired output statistics to an .sav file with OMS . If the imputed data sets had been saved to different files, then a Python script or macro should be possible to automate the process. For step 3, Example 31 in the AMOS Guide provides formulae for combining the output from several regressions to calculate a standard error for a parameter estimate, as well as links to online calculators for the CIs for estimates from multiple imputed data sets. The formulae in Ex 31 could be performed with a combination of aggregate and transformation commands in SPSS, on first inspection.

4. Mplus

The following information was obtained from Muthen, L.K. and Muthen, B.O. (1998-2007). Mplus User’s Guide. Fourth Edition. Los Angeles, CA: Muthen & Muthen. Retrieved from http://www.statmodel.com/ugexcerpts.shtml.

EXAMPLE 12.13: ANALYZING MULTIPLE IMPUTATION DATA SETS

```
TITLE: this is an example of a CFA with continuous factor indicators using
multiple imputation data sets
DATA: FILE IS impute.dat;
TYPE = IMPUTATION;
VARIABLE: NAMES ARE yl-y6;
MODEL: f1 BY yl-y3;
       f2 BY y4-y6;
```

The example above is based on Example 5.1 in which a single data set is analyzed. In this example, data sets generated using multiple imputation are analyzed. The FILE option of the DATA command is used to give the name of the file that contains the names of the multiple imputation data sets to be analyzed. The file named using the FILE option of the DATA command must contain a list of the names of the multiple imputation data sets to be analyzed. This file must be created by the user. Each record of the file must contain one data set name. For example, if five data sets are being analyzed, the contents of impute.dat would be:

```
data1.dat
data2.dat
data3.dat
data4.dat
data5.dat
```

where datal.dat, data2.dat, data3.dat, data4.dat, and data5.dat are the names of the five data sets created using multiple imputation.

When TYPE=IMPUTATION is specified, an analysis is carried out for each data set in the file named using the FILE option. Parameter estimates are averaged over the set of analyses, and standard errors are computed using the average of the standard errors over the set of analyses and the between analysis parameter estimate variation (Schafer, 1997).

The Output Command

The STANDARDIZED option is now available with the TYPE=IMPUTATION option of the DATA command.

5. HLM

The following information was obtained from Raudenbush, Stephen W., Bryk, Anthony S., Cheong, Yuk Fai, & Congdon, Richard. HLM5: Hierarchical Linear & Nonlinear Modeling. Scientific Software International, 2004.

9.2 Applying HLM to multiply-imputed data

A satisfactory solution to the missing data problem involves multiple, model-based imputation (Rubin, 1987,’Little & Rubin, 1987, Schafei, 1997). A multiple imputation procedure produces *M* “complete” data sets. Users can apply HLM2 and HLM3 to these multiply-imputed data to produce appropriate estimates that incorporate the uncertainty resulting from imputation.

There can be multiply-imputed values for the outcome or one covariate, or for the outcome and/or covariates.

HLM has two methods to analyze multiply-imputed data. They both use the same equations to compute the averages, so the method chosen depends on the data you are analyzing.

“Plausible Values” as described in Sections 9.2.1 and 9.2.3. This method is usually preferable for data sets that have only one variable (outcome or predictor) for which you have several plausible values. In this case, you need to make one MDM file containing *all* of the plausible values, plus any other variables of interest.

“Multiple imputation” as described in Section 9.2.4. This method is necessary if you have more than one variable for which you have multiply-imputed data. This method also requires a different way of setting up MDM files. Here, you have to create as many MDMs as you have plausible vales. When making these MDMs, you should use the same level-2 file (and level-3 file if using HLM3), but several level-1 files are needed.

Those variables that are not multiply imputed should be the same in all these level-1 files. The variables that *are* multiply imputed should be separated into the separate level-1 files, but they *must* have the same variable names across these level-l files, since the same model is run on each of these MDMs.

9.2.1 Data with multiply-imputed values for the outcome or one covariate

HLM2 and HLM3 enable users to produce correct HLM estimates when using data sets that contain two or more values or plausible values for the outcome variable or one covariate. One such data set is the National Assessment of Educational Progress (NAEP), an U.S. Department of Education achievement test given to a national sample of fourth, eighth, and twelfth graders.

Due to the use of balanced incomplete block (BIB) spiraling in the administration of the NAEP assessment battery, special procedures and calculations are necessary when estimating any population parameters and their standard errors with data sets such as NAEP. Every student was not tested on the same items, so item response theory (IRT) was used to estimate proficiency scores for each individual student. This procedure estimated a range or distribution of plausible values for each student's proficiency rather than an individual observed score. NAEP drew five plausible values at random from the conditional distribution of proficiency scores for each student. The measurement error is due to the fact that these scores are estimated, rather than observed.

In general, these plausible values are used to produce parameter estimates in the following way:

- Each parameter is estimated for each of the five plausible values, and the five estimates are averaged.
- Then, the standard error for this average estimate is calculated using the approach recommended by Little & Schenker (1995).
- This formula essentially combines the average of the sampling error from the five estimates with the variance between the five estimates multiplied with a factor related to the plausible values. The result is the measurement error.

In an HLM analysis, with either two- or three-levels, the parameter estimates are based on the average parameter estimates from separate HLM analyses of the five plausible values. That is, a separate HLM analysis is conducted on each of the five plausible values.

Without HLM, these procedures could be performed by producing HLM estimates for each plausible value, and then averaging the estimates and calculating the standard errors using another compute program. These procedures are tedious and time-consuming, especially when performed on many models, grades, and dependent variables.

HLM takes the plausible values into account in generating the HLM estimates. For each HLM model, the program runs each of the five (or the number specified) plausible values internally, and produces their average value and the correct standard errors. There will seem to be one estimate, but the five HLM estimates from the five plausible values are produced and their average and measurement error calculated correctly, thus ensuring an accurate treatment of plausible value data. The output is similar to the standard HLM program output, except that all the components are averaged over estimates derived from the five plausible values. In addition, the output from the five plausible value runs is available in a separate output file.

9.2.2 Calculations performed

The program conducts a separate HLM analysis for each plausible value. The output of the separate HLM analyses is written to files with consecutive numbers, for example, OUT.1, OUT.2, OUT.3, etc. Then, HLM calculates the average of the parameter estimates from the separate analyses and computes the standard errors. The output of the average HLM parameter estimates and their standard errors is found in the output file with the extension AVG.

9.2.2.1 Average parameter estimates

The following parameter estimates are averaged by HLM:

- The fixed effects (gammas)
- The reliabilities
- The parameter variances (tau) and its correlations
- The chi-square values to test whether the parameter variance is zero
- The standard errors for the variance-covariance components (full maximum likelihood estimates)
- Multivariate hypothesis testing for fixed effects

9.2.2.2 Standard error of the gammas

The standard error of the averaged fixed effects (gammas) is estimated as described below. The Student's t-value is calculated by dividing the average gamma by its standard error, and the probability of the t-value is estimated from a standard t-distribution table.

The standard error of the gammas consists of two components — sampling error and measurement error. The following routine provided in the NAEP *Data Files User Guide* (Rogers, *et al.,* 1992) is used to approximate the component of error variance due to the error in imputations and to add it to the sampling error.

Let θ_m, *(m* **=** 1,..., *M)* represent the *m-th* plausible value. Let t_m represent the parameter estimate based on the m-th plausible value. Let U_m represent the estimated variance of t_m

- Five HLM runs were conducted based on each plausible value ***m.*** The parameter estimates from these runs were averaged:

$$t^* = \frac{\sum_{m=1}^{M} \hat{t}_m}{M}$$

- The variances of the parameters from these runs were averaged:

$$U^* = \frac{\sum_{m=1}^{M} U_m}{M}$$

- The variance of the m estimates, $t_{m,}$ was estimated:

$$B_m = \frac{\sum_{m=1}^{M}\left(\hat{t}_m - t^*\right)}{(M-1)}$$

- The final estimate of the variance of the parameter estimate is the sum of the two components:

  $V = U^* + \left(1 + M^{-1}\right) B_m$

  where the degrees of freedom is compute:

  $d.f. = (M-1)(1+r)^2,$

  where

  $$r = \frac{1+U^*}{B\left(1+\frac{1}{M}\right)}.$$

The square root of this variance is the standard error of the gammas, and it is used in a standard Student's t formula to evaluate the statistical significance of each gamma.

9.2.4 Data with multiply-imputed values for the outcome and covariates

There may be multiply-imputed values for both the outcome and the covariates. To apply HLM to such data, it is necessary to prepare as many MDM files as the number of imputed data sets. Thus, if there are five imputed data sets, five MDM files with identical variable labels need to be prepared. To run these models in batch mode, refer to Section E.3 in Appendix E.

Below are the commands for running an analysis with multiply-imputed data sets via Windows mode.
1. After specifying the model, select the **Estimation Settings** option from the **Other Settings** menu.
2. Choose **Multiple Imputation** to open the **Multiple Imputation MDM** files dialog box (See Figure 9.6 for an example).
3. Enter the names of the MDM files that contain the multiply-imputed data either by typing into the **File** # edit boxes or clicking **Browse** to open them.
4. Click **OK.** Model specification follows the usual format.

The calculations involved are very similar to the ones mentioned in Section 9.2.2.

E.3 Commands to apply HLM to multiply-imputed data

To analyze data with multiply-imputed values for the outcome variable or only one covariate, the user needs to manually add the following line into the command file:

```
        PLAUSVALS: VARLIST
```

where VARLIST lists variables containing the multiply-imputed values.

To analyze data with multiply-imputed values for the outcome and/or covariates, the user needs to prepare multiple MDM files. After setting up the multiple MDM files, the user has to submit the command files to HLM2 and HLM3 as many times as the number of multiple MDM files with an extra flag, -MI#, where # is the sequence number, starting from 0. On the last run, you also need the -E flag, (E for estimate).

Suppose there are 4 sets of multiply-imputed data for a two-level model, called MDATA1.MDM, MDATA2.MDM, MDATA3.MDM, and MDATA4.MDM and the command file is ANALYSE. MLM; the following commands need to be typed in at the system prompt:

```
        HLM2 –MI0 MDATA1.MDM  ANALYSE.MLM
        HLM2 –MI1 MDATA1.MDM ANALYSE.MLM
        HLM2 –MI2 MDATA2.MDM ANALYSE.MLM
        HLM2 –MI3 –E MDATA.4MDM ANALYSE.MLM
```