MULTIPLE IMPUTATION

Adrienne D. Woods

Methods Hour Brown Bag

April 14, 2017

A COLLECTIVIST APPROACH TO BEST PRACTICES

- As I began learning about MI last semester, I realized that there are a lot of guidelines that are not often followed…
- …or, if they are, nobody reports what they did!
- …or, guidelines that are outdated and/or different across disciplines

- This talk is…
  - Focused primarily on large samples (ECLS-K ~21,400)…
  - …on issues associated with MNAR data
  - …in the hopes of sharing what I've learned (and mitigating future frustration)
  - …Open to debate/discussion!

THE WHY: MISSING DATA!

DISCUSS: Why might you choose to impute data?

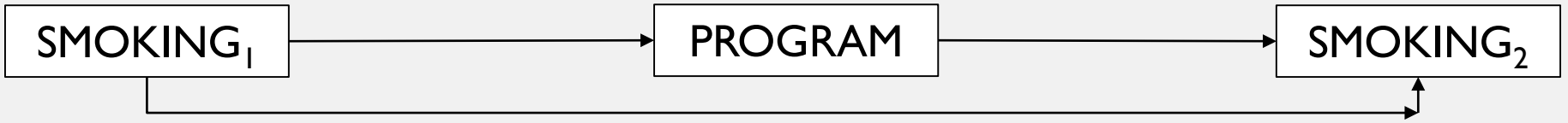
- Most commonly, folks impute due to issues of power associated with reduced sample size
  - Several methods of dealing with missing data…but also, several less efficient/poorer alternatives than MI (i.e., mean substitution)
  - "Missing by design" studies

THE WHY: TYPES OF MISSING DATA

- Missing Completely at Random
- Missing at Random
- Missing Not at Random
  - ***DISCUSS***: How do you define this?

THE WHY: TYPES OF MISSING DATA

- Missing Not at Random
  - Graham (2009): "non-ignorable missingingess"



  - Tabachnick & Fidell (2013): MNAR is related to the DV, as determined by *significant t-tests with the DV*
    - η^2 for effect sizes in large samples

THE WHY: TYPES OF MISSING DATA

- Missing Not at Random
  - Issue: no way to truly determine MAR vs. MNAR in your data

"[Controlling] variables that help account for the mechanisms resulting in missing data (e.g., race/ethnicity, age, gender, SES)…leads to a reasonable assumption of missing at random (MAR)." Hibel, Farkas, & Morgan, 2010

Is this good enough?

Even if researchers have MNAR data, they typically still impute…

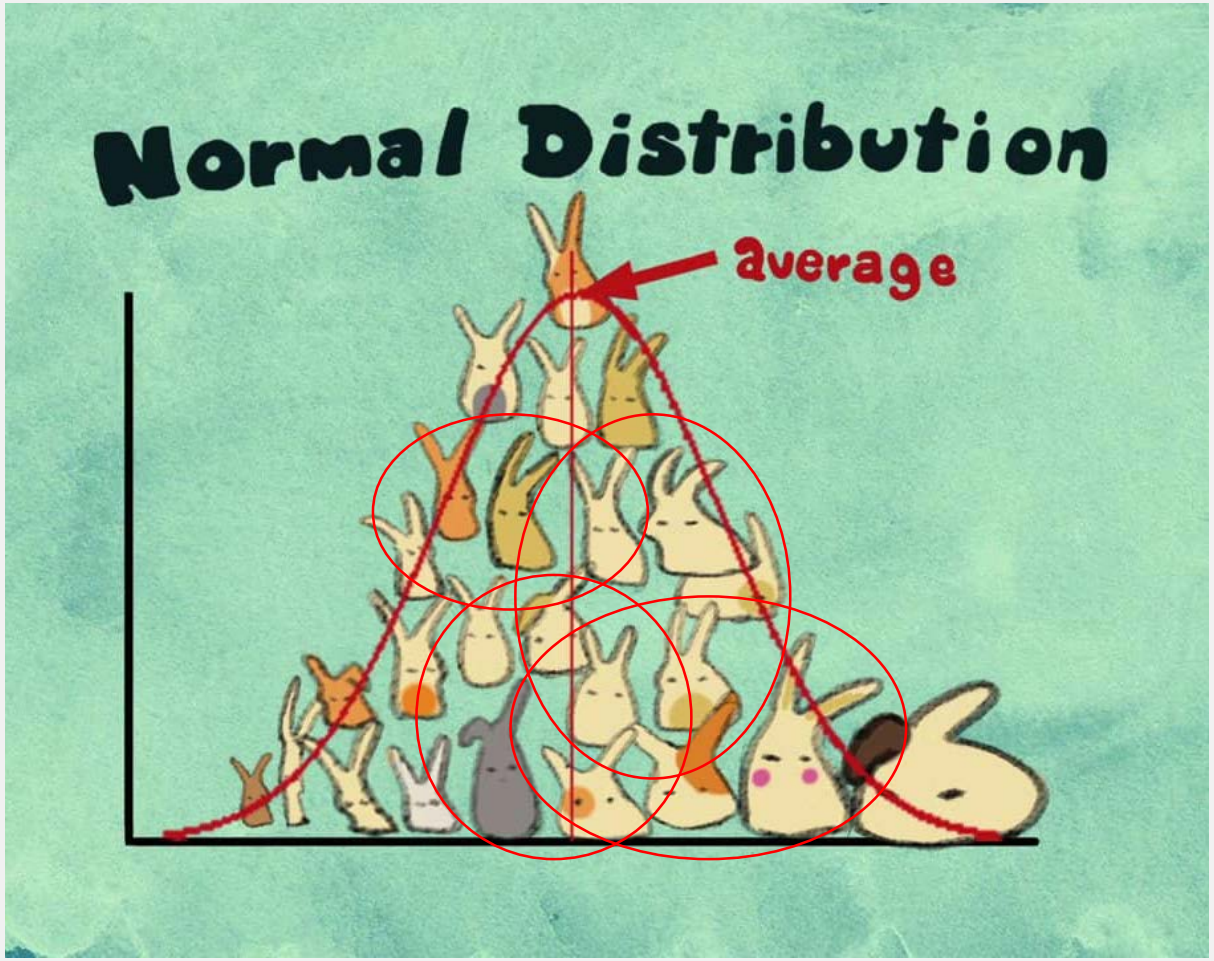
- T&F (2013) recommend modeling predictors of missingness alongside other variables as dummies
- In small samples with nonnormality, MI performed similarly to FIML (Shin, Davison, & Long, 2016)
- But, *estimates will still be biased!*

THE WHAT: WHAT IS MULTIPLE IMPUTATION?

"To the uninitiated, multiple imputation is a bewildering technique that differs substantially from conventional statistical approaches. As a result, the first-time user may get lost in a labyrinth of imputation models, missing data mechanisms, multiple versions of the data, pooling, and so on."

– Van Buuren & Groothuis-Oudshoorn (2011)

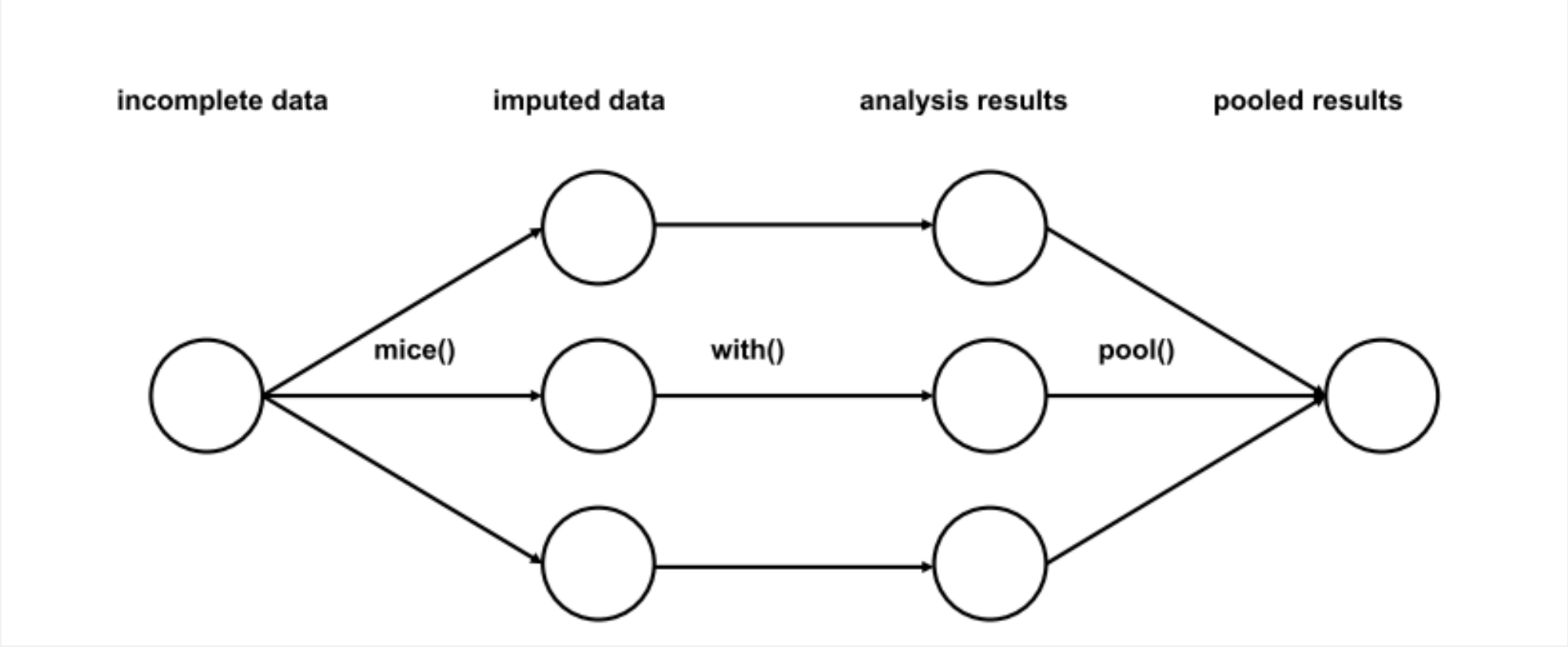
THE WHAT: WHAT IS MULTIPLE IMPUTATION?



- Single imputation methods (mean replacement, regression, etc.) assume perfect estimation of imputed values and ignore between-imputation variability
  - May result in artificially small standard errors and increased likelihood of Type I errors, and are only appropriate for MCAR data
  - Imputed values from single imputation always lie right on the regression line; but, real data always deviate from the regression line by some amount
- MI creates several datasets with estimated values for missing information
  - Incorporates uncertainty into the standard errors of imputed values by accounting for variability between imputed solutions

Acock, 2005; Graham, 2009; Hibel, Farkas, & Morgan, 2010; Schafer, 1999

THE WHAT: WHAT IS MULTIPLE IMPUTATION?



Van Buuren & Groothuis-Oudshoorn (2011): Seven Choices

BEFORE VS. AFTER MI

Table III. UK700 data: data format before and after imputation, for 5 selected individuals and 2 imputed data sets. The imputed data set includes two added identifiers, `_mi` for individuals and `_mj` for imputed data sets, and includes the original data as `_mj==0`.

Before imputation		
sat94	rand	sat96
20	0	.
18	1	22
17	0	16
.	1	.
.	0	14

After imputation			
_mj	sat94	rand	sat96
0	20	0	.
0	18	1	22
0	17	0	16
0	.	1	.
0	.	0	14
1	20	0	8.35
1	18	1	22
1	17	0	16
1	20.28	1	13.66
1	19.53	0	14
2	20	0	11.10
2	18	1	22
2	17	0	16
2	24.91	1	16.81
2	22.76	0	14

THE WHAT: WHAT IS MULTIPLE IMPUTATION?

growth curve analyses were conducted in LISREL 8.53 using Maximum Likelihood estimation. Missing data was handled using multiple imputation with an Expectation Maximization algorithm.

were complete data on the scholar-level variables selected. To retain the largest number of cases in the analytic sample, we used multiple imputation to estimate missing values for four student-level variables for which complete data did not exist. Table 1 provides the general characteristics of the analytic sample relative to the full sample.

We used multiple imputation to account for missing data in both the ECLS-B and ECLS-K. Specifically, and for each analytical data set, we imputed 5 (complete) data sets, estimating models separately for each completed data set, and then combining these estimates into a single set of estimates using mathematically derived formulas (Little & Rubin, 2002). Only observations with missing independent (predictor) variables had those values imputed; cases with missing outcome variables were deleted.

dence of good fit. The chi-square model fit test is also reported. There was very little missing data for this study, ranging from 0.3% to 6.9% for all the variables. However, to increase power for the predictor variables, we conducted multiple imputation for missing values using Bayesian estimation through Mplus. For the outcome variables, methods robust to missing data were utilized for data analysis (i.e., no missing values were deleted).

intervention and control group for the number of hours in special education. Multiple imputation was used to impute missing values for number of hours in special education/resource (Rose and Fraser 2008). Multiple imputation (10 in each group) was conducted separately for each intervention group using SAS (V 9.1). Ten imputations are adequate for most applications if values are missing at random. (Acock 2005). Pooled estimates of the parameters and standard errors from the combined imputed data set were used.

cally significant subgroup effects across all band-widths are available upon request.

In fitting all our regression models, we used the method of multiple imputation (with 50 imputations) to account for missing data, following Graham (2009). In Table 1, we present summary statistics on the child outcomes, including the percent missing for each outcome.

rice, Laird, & Ware, 2004, chap. 14).

Missing data were imputed 20 times using PROC MI in SAS version 8.1 with the recommended expectation-maximization algorithm and Markov chain Monte Carlo method (Schafer & Graham, 2002). Analyses were run on each of the 20 data sets with results combined according to Rubin's rules (Rubin, 1987) using PROC MIANALYZE. The pattern of significant findings did not differ between analyses based on imputed data and results of the same analyses performed on the original, nonimputed data using listwise deletion procedures.

Van Buuren & Groothuis-Oudshoorn (2011): Seven Choices

THE HOW: GUIDELINES FOR MI

1. Decide whether data are MAR or MNAR – latter requires additional modeling assumptions
2. Form of imputation model
   - Depends on scale of each variable to be imputed
   - Incorporates knowledge about relationship between variables

Method	Description	Scale type
`pmm`	Predictive mean matching	numeric
`norm`	Bayesian linear regression	numeric
`norm.nob`	Linear regression, non-Bayesian	numeric
`mean`	Unconditional mean imputation	numeric
`2L.norm`	Two-level linear model	numeric
`logreg`	Logistic regression	factor, 2 levels
`polyreg`	Multinomial logit model	factor, >2 levels
`polr`	Ordered logit model	ordered, >2 levels
`lda`	Linear discriminant analysis	factor
`sample`	Random sample from the observed data	any

THE HOW: GUIDELINES FOR MI

3. Which variables should you include as predictors in the imputation model?
   - Any variables you plan to use in later analyses (including controls)
   - General advice: use as many as possible (could get unwieldy!)
   - Although, some (i.e., Kline, 2005; Hardt, Herke, & Leonhart, 2012) believe that this introduces more imprecision, especially if the auxiliary variable explains less than 10% of the variance in missingness on Y… ***thoughts?***

AN EXAMPLE…

	Math Competency		School Belongingness	
	Attempt 1	Attempt 2	Attempt 1	Attempt 2
	Std. B (SE)	Std. B (SE)	Std. B (SE)	Std. B (SE)
Constant	0.54 (.61)	1.39 (.75)	1.97 (.43)***	2.08 (.54)***
Male	0.06 (.06)	0.05 (.06)	-0.04 (.04)	-0.04 (.04)
Black	0.23 (.09)**	0.13 (.07)	-0.10 (.06)	-0.05 (.05)
Hispanic	0.04 (.07)	0.03 (.07)	-0.08 (.05)	-0.05 (.05)
Asian	-0.06 (.15)	-0.01 (.14)	0.02 (.10)	0.02 (.09)
K-8 Read Gain	-0.22 (.15)	-0.22 (.13)	-0.01 (.10)	0.08 (.10)
K-8 Math Gain	0.83 (.17)***	0.78 (.16)***	0.09 (.02)	0.07 (.11)
Special Ed. Dosage	0.08 (.03)**	0.07 (.03)*	0.04 (.02) +	0.05 (.02)*
Special Ed. Recency	0.01 (.03)	0.02 (.02)	-0.01 (.02)	-0.01 (.02)

$+p < .10$, $*p < .05$, $**p < .01$, $**p < .001$

What I changed:

- Accidentally left out three variables that I wanted to use in my analysis model as autoregressive controls (bolded)
- Both $m = 70$
- Predictors of interest are *Special Ed. Dosage* and *Special Ed. Recency* (did not impute into the latter)

Stata Code (second attempt)

mi impute chained (pmm, knn(10)) R1_KAGE WKSESL WKMOMED C7SDQRDC C7SDQMTC C7SDQINT C7LOCUS C7CONCPT belong peers C1R4RSCL C1R4MSCL readgain mathgain C5SDQRDC C5SDQMTC C5SDQINT C6SDQRDC C6SDQMTC C6SDQINT **C5SDQPRC C6SDQPRC** T1LEARN T1CONTRO T1INTERP T1INTERN T1EXTERN **P1NUMSIB** (logit) youngma retained single headst premat (ologit) C7HOWFAR C7LONLY C7SAD sped_dos = sped_rec race_r gender, add(1) rseed(53421) burnin(100) dots force augment

THE HOW: GUIDELINES FOR MI

4. Imputing variables that are functions of other (incomplete) variables
    - Sum scores, interaction variables, ratios, etc…
    - DON'T transform! (could impute outliers; Graham, 2009)
    - Standardized variables??? (my guess is no…)
5. Order in which variables should be imputed
6. Setup of starting imputations and the number of iterations
    - Includes *k*-nearest neighbors if using predictive mean matching

THE HOW: GUIDELINES FOR MI

7. How many multiply imputed datasets, *m*, should you create?

- Previously, *m* = 3-5 considered acceptable in social sciences
- But, your estimates can change, especially if you have MNAR data…

  i.e., in *m* = 3, *p* = .04… in *m* = 10, *p* = .08

- "Impute one dataset, see how long it takes, and then base your decision about *m* on time constraints and software capability." (Van Buuren & Groothuis-Oudshoorn, 2011)

  NO. 😡

New rule: more is better!

- "Setting *m* too low may result in large simulation error, especially if the ***fraction of missing information*** is high."

THE HOW: GUIDELINES FOR MI

- ***Fraction of Missing Information*** (FMI)
  - Statistical formula based on the amount of missing data in the simplest case (Rubin, 1987)
  - Rule of thumb: set *m* equal to the number of incomplete cases, which will typically be less than the FMI
  - Relative efficiency of imputations: FMI/m ~= .01
  - Annoying in that this depends on *m,* but *m* depends on FMI (Spratt et al., 2010)
    - But, you could impute a few datasets, check FMI, then impute again…then check FMI again! (White, Royston, Wood, 2011; Graham, Olchowski, & Gilreath, 2007)

AN EXAMPLE…

First, imputed one dataset to make sure the code worked without error. Then, imputed up to m = 4 to check FMI:

```
Multiple-imputation estimates                 Imputations       =          4
Multinomial logistic regression               Number of obs     =      4,359
                                              Average RVI       =     0.2141
                                              Largest FMI       =     0.6596
DF adjustment:   Large sample                 DF:     min       =       8.65
                                                      avg       = 143,247.46
                                                      max       =   1.94e+07
Model F test:       Equal FMI                 F( 165,15025.7)   =       4.43
Within VCE type:        Robust                Prob > F          =     0.0000
```

FMI/*m* = 0.6596/4 = .165

Then, imputed another 46 datasets to get to m = 50, and checked FMI again:

```
Multiple-imputation estimates                 Imputations       =         50
Multinomial logistic regression               Number of obs     =      4,359
                                              Average RVI       =     0.1927
                                              Largest FMI       =     0.3521
DF adjustment:   Large sample                 DF:     min       =     402.64
                                                      avg       =  28,528.17
                                                      max       = 813,522.80
Model F test:       Equal FMI                 F( 145,259060.3)  =       4.81
Within VCE type:        Robust                Prob > F          =     0.0000
```

FMI/*m* = 0.3521/50 = .007

SOFTWARE PACKAGES

- R – mice package
  - Completely syntax-based, can get out of hand for uninitiated/beginners
- STATA – multiple imputation feature
  - Subsequent data analyses conducted with "mi estimate:" as the precursor to code
- SPSS – multiple imputation feature
  - Creates one dataset or imputes X separate datasets (useful for HLM, for example)
  - But, limited in options
    - e.g., can't manipulate knn

CO-CONSTRUCTED KNOWLEDGE & DISCUSSION:

Main Take-Aways:

- First, always know what type of missing data you are working with
- Base *m* on FMI – rule of thumb is FMI/*m* < .01
- Know your analysis model beforehand and include *at least* all analysis variables in imputation model (including interaction terms)
- Above all, ***be explicit about your choices***.
  - Include software you used to impute, auxiliary variables, etc.
  - If not written out in actual manuscript, add to appendices!

…Other discussion points or best practices?

…What might be some alternatives to multiple imputation that folks could use, and why?

THANK YOU! ☺

REFERENCES

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549-576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206-213.

Rubin, D. B. (1987). Comment. *Journal of the American Statistical Association*, *82*(398), 543-546.

Shin, T., Davison, M. L., & Long, J. D. (2016). Maximum Likelihood Versus Multiple Imputation for Missing Data in Small Longitudinal Samples With Nonnormality.

Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *American journal of epidemiology*, *172*(4), 478-487.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th Ed.). Pearson.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*(3).

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377-399.

RELATIVE EFFICIENCY OF *M*

"The variability between sets of imputations depends on both the number of imputations used and the fraction of missing information. However, ***the fraction of missing information*** is itself estimated using the between- and within-imputation variances, and thus may have substantial variability when estimated from small numbers of imputations. Monte Carlo variation among sets of small numbers of imputations can be substantial enough to materially affect conclusions, particularly where the original data set is small. One approach might be to estimate the Monte Carlo variation and use that to decide the appropriate number of imputations." (p. 486, Spratt et al., 2010)

"The early literature focused on efficiency, and the conclusion was that you could usually get by with three to five data sets. Schafer (1999) upped that number slightly when he stated that "Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations." That conclusion was based on Rubin's formula for relative efficiency: 1/(1+F/M) where F is the ***fraction of missing information*** and M is the number of imputations. Thus, even with 50% missing information, five imputed data sets would produce point estimates that were 91% as efficient as those based on an infinite number of imputations. Ten data sets would yield 95% efficiency. But what's good enough for efficiency isn't necessarily good enough for standard error estimates, confidence intervals, and p-values." (Allison, 2012)