Prev Sci (2007) 8:206–213
DOI 10.1007/s11121-007-0070-9

How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory

John W. Graham · Allison E. Olchowski · Tamika D. Gilreath

Published online: 5 June 2007

Abstract Multiple imputation (MI) and full information maximum likelihood (FIML) are the two most common approaches to missing data analysis. In theory, MI and FIML are equivalent when identical models are tested using the same variables, and when *m*, the number of imputations performed with MI, approaches infinity. However, it is important to know how many imputations are necessary before MI and FIML are sufficiently equivalent in ways that are important to prevention scientists. MI theory suggests that small values of *m*, even on the order of three to five imputations, yield excellent results. Previous guidelines for sufficient *m* are based on relative efficiency, which involves the fraction of missing information (γ) for the parameter being estimated, and *m*. In the present study, we used a Monte Carlo simulation to test MI models across several scenarios in which γ and *m* were varied. Standard errors and p-values for the regression coefficient of interest varied as a function of *m*, but not at the same rate as relative efficiency. Most importantly, statistical power for small effect sizes diminished as *m* became smaller, and the rate of this power falloff was much greater than predicted by changes in relative efficiency. Based our findings, we recommend that researchers using MI should perform many more imputations than previously considered sufficient. These recommendations are based on γ, and take into consideration one's tolerance for a preventable power falloff (compared to FIML) due to using too few imputations.

J. W. Graham (✉) · A. E. Olchowski · T. D. Gilreath
Department of Biobehavioral Health, Penn State University,
E-315 Health & Human Development Bldg.,
University Park, PA 16802, USA
e-mail: jgraham@psu.edu

Keywords Multiple imputation · Number of imputations · Full information maximum likelihood · Missing data · Statistical power

Since Rubin's (1987) classic book on the subject, multiple imputation has enjoyed a steady growth in popularity and usefulness. Technical articles, books, and multiple imputation software abound (e.g., Collins et al. 2001; Graham et al. 2003; King et al. 2001; Schafer, 1997; Schafer and Graham 2002; Schafer and Olsen 1998). Perhaps a more telling indication of the value of the procedure is the plethora of substantive articles and chapters that make use of multiple imputation (for example, http://www.multiple-imputation.com/ lists 440 multiple-imputation-related publications as of May 2006).

The main idea of multiple imputation is that plausible values may be used in place of the missing values in a way that allows (1) parameter estimates to be unbiased, and perhaps more important, (2) the uncertainty of parameter estimation in the missing data case to be estimated in a reasonable way. This ability to estimate the uncertainty of parameter estimation in the missing data case is due to what is often referred to as "Rubin's rules" for combining the results of analysis of multiply imputed datasets (Rubin 1987). The point estimate of each parameter (e.g., a regression coefficient, *b*) is simply the average of the parameter estimate $(\overline{b})$ obtained over the *m* imputed datasets. But it is the standard error for the parameter estimate that really makes multiple imputation a uniquely useful tool. In multiple imputation, the variance of estimation is partitioned into the within imputation variance, which captures the usual kind of sampling variability, and the between imputation variance, which captures the estimation variability due to

missing data. Formulas for these quantities, adapted from Schafer (1997) are:

$$U_b = \sum SE_b^2 \Big/ m$$

for the within imputation variance of, say, a particular regression coefficient, where U_b is the average of the squared standard error (*SE*) for that regression coefficient over the m imputed datasets, and

$$B_b = 1 \Big/ (m-1) \sum \left(b - \overline{b}\right)^2$$

for the between imputation variance. B_b is the sample variance of the parameter estimate over the m imputed datasets. The formula for combining these two variances, also adapted from Schafer (1997), is

$$T_b = U_b + [1 + (1/m)]B_b$$

and

$$SE_b = sqrt(T_b)$$

The parameter estimate is then divided by its *SE* to give a *t*-value. The degrees of freedom (*df*) for this *t*-value, again adapted from Schafer (1997), is:

$$df = (m-1)[1 + (mU_b/(m+1)B_b)]^2$$

The *t*-value, along with its *df* may be used for statistical inference. If one prefers, SE_b may be used in the usual way for calculating 95% confidence intervals.

Another quantity that figures prominently in multiple imputation is known as the fraction of missing information (γ). Schafer and Olsen (1998) give the formula for γ as

$$\gamma = \frac{r + 2/(df+3)}{r+1}$$

where

$$r = \frac{(1 + m^{-1})B}{U}$$

Although γ is the same as the amount of missing data in the simplest case, it is typically rather different from (less than) the amount of missing data, per se, in more complicated situations (Rubin 1987; p. 114). For example, if other variables included in the imputation model are highly correlated with the (sometimes missing) variables of interest, then the amount of missing information is generally smaller than the percentage of missing data.

How Many Imputations are Needed: Previous Thinking

An important aspect of previous technical treatments of multiple imputation (e.g., Rubin 1987; Schafer 1997; Schafer and Olsen 1998) is the discussion of the number of imputations that are needed for good statistical inference. For example, Schafer and Olsen (1998) suggest the following.

> In many applications, just 3–5 imputations are sufficient to obtain excellent results. ... Many are surprised by the claim that only 3–5 imputations may be needed. Rubin (1987, p. 114) shows that the efficiency of an estimate based on m imputations is approximately
>
> $$\left(1 + \frac{\gamma}{m}\right)^{-1},$$
>
> where γ is the fraction of missing information for the quantity being estimated.... gains rapidly diminish after the first few imputations. ... In most situations there is simply little advantage to producing and analyzing more than a few imputed datasets (pp. 548–549).

Meaning of "Efficiency"

What does it mean to say that the "efficiency of the estimate" is given by $(1 + \gamma/m)^{-1}$? "Efficiency", a quantity that is very common in statistics, is based on the mean-square error (MSE) for one estimator compared to another. In this case, we could calculate the MSE, or the mean of the squared error, as:

$$MSE = (b - \beta)^2 \Big/ N$$

where b is the estimated regression coefficient, and β is the population value of that regression coefficient. N in this case might be the number of random draws from the population or the number of replications of a simulation.

Missing Data Methods: FIML vs. MI

Missing data theorists have argued that MI and FIML are equivalent in theory, but not as practiced. Collins et al. (2001) showed the value of including auxiliary variables (variables not part of the model under study) in the missing data model. It is an easy matter to include auxiliary variables with MI, but FIML users rarely do so. Graham's (2003) models allow one to incorporate auxiliary variables into FIML-based SEM models without altering the meaning of the substantive model under study, thereby making it easier for FIML users to make their analyses equivalent to MI in this important sense.

Another way to compare equivalence of MI and FIML involves the number of imputations (m) used with MI. We take it as an axiom that MI and FIML are equivalent when

the variables and models tested are the same, and when $m=\infty$. But what m is needed to approximate $m=\infty$? As noted above, MI theorists have argued that surprisingly small m is needed for efficient estimation. Unfortunately, relative efficiency is a quantity with little practical meaning for prevention scientists. And, as we demonstrate in this article, γ itself is unreliably estimated unless m is rather large. Because one's best choices of missing data analysis in most cases are MI and FIML, it will be important to know for what m MI is truly equivalent to FIML.

In this article, we expand on what one actually gets with fewer or more imputations. We conduct a brief Monte Carlo simulation to demonstrate our main points. We demonstrate that the empirical estimates of efficiency, as defined above, are rather close to the theoretical predictions given by Schafer and Olsen (1998). However, we also show that other important quantities, such as standard errors of the estimate, p-values, and power all vary rather markedly with the number of imputations (m). In particular, we show that one of these quantities, statistical power can vary rather more dramatically with m than is implied by the efficiency tables presented in previous discussions of MI theory. Furthermore, we evaluate the equivalence of MI and FIML across multiple data scenarios involving variable levels of γ.

Materials and Methods

A Monte Carlo Simulation

For our simulation, we first generated 100,000 cases for two normally distributed variables, X and Y (data were generated using Jöreskog & Sörbom's utility GENRAW.) In this population, the regression coefficient for X predicting Y was β=.0969. Second, for each replication of the simulation, some number of cases were drawn at random from the population, as shown in Table 1, depending on the value of γ (within each replication, elements were drawn from the population without replacement; however, the same element could be drawn for two or more replications). The values for Y for all but 800 of those cases were set to missing (completely at random). That is, for each level of γ, the number of complete cases was held constant at 800. As γ increased, the proportion of cases with missing data relative to those with complete data increased.

Table 1 Simulation sample sizes drawn from the population

γ	N selected from Population
0.10	889
0.30	1,143
0.50	1,600
0.70	2,667
0.90	8,000

For each level of γ N=800 cases had no missing data.

Third, the missing values were imputed using m=3, 5, 10, 20, 40, or 100 imputations (SAS Proc MI, versions 8.2 and 9.1, was used for the simulation). Fourth, a simple regression analysis (PROC REG) was performed on the resulting datasets (X predicting Y), and the results were saved. In total, there were five levels of γ (.1, .3, .5, .7, .9) and six levels of m, yielding 30 cells for the simulation. We used 8000 replications for each of these 30 cells.

Results

The main results of the simulation are presented in Table 2. The first thing to note in Table 2 is that the regression coefficients were essentially unbiased for all values of m and all values of γ. Then, within each level of γ, as the number of imputations decreased from m=100 to m=3: (1) the values of MSE and SE increased; (2) power (the probability of rejecting the false null hypothesis) was reduced (for γ=0.5, for example, this reduction was from .78 to .59); (3) the estimate of γ differed somewhat more from its true value; and (4) the variability of the estimate of γ increased as m decreased; this increase in variability was highest for intermediate values of γ.

Table 3 rearranges some of the key findings of Table 2 and provides a direct comparison with values calculated from the efficiency formula from MI theory. Column 9 (labeled "Relative Efficiency: MI Theory") shows the efficiency based on Schafer and Olsen's (1998) formula for a particular m compared to m=100 for that same level of γ. Column 8 (labeled "Relative Efficiency: Empirical") shows the same values derived from our simulation. These two columns are not the same, of course, but in terms of absolute values, these two columns are more similar to each other than they are to any other column in this table. That is, despite the slight simulation "wobble", our simulated estimates of efficiency map rather well onto the theoretical values derived from Rubin's formula.

Columns 5, 6, and 7 (located under the heading "Percent of Optimal") show what happens to statistical power, SE, and the p value as the number of imputations decrease from m=100 to m=3. These figures are presented in a metric that allows a direct comparison with the "Relative Efficiency: MI Theory" values (column 9). Column 6 (labeled "SE") shows the m=100 SE value divided by the each of the remaining SE values. Note that the deviations from the optimal SE (i.e., SE for m=100), based on the simulation results, are much less dramatic than the falloff in efficiency implied by MI theory (column 9). Column 2 (labeled "Power") is taken from Table 2.

Table 2 Results of Monte Carlo simulation

m	Power	*b*	*SE*	*t*	*df*	*p*	γ	SDγ	MSE ($\times 10^3$)
(γ=0.10)									
100	0.7910	0.0972	0.0353	2.76	10.8 K	0.049	0.101	0.0194	1.2022
40	0.7880	0.0969	0.0353	2.75	4,527	0.050	0.102	0.0250	1.2270
20	0.7846	0.0972	0.0353	2.76	2,454	0.050	0.105	0.0332	1.2203
10	0.7799	0.0968	0.0354	2.74	1,711	0.052	0.109	0.0483	1.2429
5	0.7760	0.0968	0.0355	2.73	4,714	0.052	0.119	0.0766	1.2288
3	0.7620	0.0967	0.0357	2.72	5,562 K	0.056	0.131	0.1143	1.2967
(γ=0.30)									
100	0.7881	0.0969	0.0353	2.75	1137	0.048	0.303	0.0353	1.1954
40	0.7873	0.0974	0.0353	2.77	471	0.049	0.306	0.0524	1.2120
20	0.7824	0.0975	0.0355	2.76	249	0.051	0.311	0.0726	1.2339
10	0.7613	0.0963	0.0356	2.72	157	0.056	0.320	0.1064	1.2346
5	0.7308	0.0965	0.0360	2.72	370	0.062	0.337	0.1611	1.2880
3	0.6873	0.0971	0.0364	2.75	173 K	0.071	0.348	0.2215	1.3106
(γ=0.50)									
100	0.7809	0.0965	0.0353	2.74	403	0.051	0.503	0.0399	1.2247
40	0.7763	0.0970	0.0354	2.75	164	0.052	0.506	0.0596	1.2390
20	0.7719	0.0978	0.0356	2.77	84	0.053	0.512	0.0851	1.2494
10	0.7479	0.0974	0.0359	2.76	47	0.059	0.521	0.1243	1.2709
5	0.6819	0.0967	0.0361	2.76	62	0.071	0.525	0.1840	1.3545
3	0.5863	0.0972	0.0368	2.80	48 K	.093	.523	.2543	1.4361
(γ=0.70)									
100	0.7780	0.0971	0.0354	2.75	203	0.052	0.703	0.0327	1.2451
40	0.7710	0.0967	0.0353	2.76	82	0.055	0.704	0.0505	1.2602
20	0.7486	0.0965	0.0356	2.75	41	0.059	0.709	0.0721	1.2753
10	0.7116	0.0966	0.0358	2.77	21	0.066	0.712	0.1056	1.2878
5	0.6096	0.0969	0.0366	2.82	25	0.087	0.713	0.1660	1.3872
3	0.4930	0.0962	0.0368	2.93	1,413 K	0.118	0.688	0.2403	1.4836
(γ= 0.90)									
100	0.7756	0.0964	0.0353	2.75	122	0.053	0.901	0.0136	1.2057
40	0.7618	0.0970	0.0355	2.77	48	0.055	0.902	0.0211	1.2507
20	0.7291	0.0968	0.0356	2.78	24	0.063	0.903	0.0322	1.3216
10	0.6689	0.0967	0.0360	2.83	12	0.075	0.903	0.0520	1.3517
5	0.5334	0.0966	0.0365	2.97	6	0.102	0.895	0.0997	1.4009
3	0.3876	0.0969	0.0364	3.33	236	0.147	0.862	0.1782	1.6662

Figures for each cell were based on 8,000 replications. The population $r=b=0.0969$. Theoretical power=0.7839 for $N=800$. Power for equivalent FIML analysis was also 0.7839 (for all levels of γ).

Columns 3 and 4 (located under the heading "Power Falloff") show the power falloff when *m* is small compared to $m=100$ (column 3) and the comparable FIML analysis (column 4). Column 3 shows the percent by which each power figure is less than the power observed for $m=100$. Note that the power falloff shown by our simulation is rather more dramatic than the falloff of efficiency predicted by MI theory, especially as *m* gets small. Column 4 shows the percent by which each power figure is less than the power for the corresponding FIML model (0.7839). For $\gamma>0.30$, the falloff compared to the FIML analysis is slightly higher than that for $m=100$.

The numbers presented in Table 3 show that efficiency is a quantity that must be evaluated carefully. It is rather clear, for example, that this quantity does not reflect the actual increase in the standard error as the number of imputations is diminished. Nor does it reflect the increase in the *p* value; the *p* value increased much more rapidly than predicted by the efficiency formula as *m* goes from 100 to 3.

Details of Power Falloff

Most importantly, it is rather clear that the drop in efficiency does not reflect the loss of power seen in our

Table 3 Rearranged simulation results

		Power falloff			Percent of optimal			Relative efficiency	
	m (1)	Power (2)	m=100 (3; %)	FIML (4; %)	Power (5)	SE (6)	p value (7)	Empirical (8)	MI theory (9)
γ=.10	100	0.79		0					
	40	0.79	0.4	0	1.0	1.0	0.98	0.98	1.0
	20	0.78	0.8	0	0.99	1.0	0.98	0.99	1.0
	10	0.78	1.4	0.5	0.99	1.0	0.94	0.97	0.99
	5	0.78	1.9	1.0	0.98	0.99	0.94	0.98	0.98
	3	0.76	3.7	2.8	0.96	0.99	0.88	0.93	0.97
γ=.30	100	0.79		0					
	40	0.79	0.1	0	1.0	1.0	0.98	0.99	1.0
	20	0.78	0.7	0.2	0.99	0.99	0.94	0.97	0.99
	10	0.76	3.4	2.9	0.97	0.99	0.86	0.97	0.97
	5	0.73	7.3	6.8	0.93	0.98	0.77	0.93	0.95
	3	0.69	13	12.3	0.87	0.97	0.68	0.91	0.91
γ=.50	100	0.78		0.4					
	40	0.78	0.6	1.0	0.99	1.0	0.98	0.99	0.99
	20	0.77	1.2	1.5	0.99	0.99	0.96	0.98	0.98
	10	0.75	4.2	4.6	0.96	0.98	0.86	0.96	0.96
	5	0.68	13	13	0.87	0.98	0.72	0.90	0.91
	3	0.59	25	25	0.75	0.96	0.55	0.85	0.86
γ=.70	100	0.78		0.8					
	40	0.77	0.9	1.6	0.99	1.0	0.95	0.99	0.99
	20	0.75	3.8	4.5	0.96	0.99	0.88	0.98	0.97
	10	0.71	8.5	9.2	0.91	0.99	0.79	0.97	0.94
	5	0.61	22	22	0.78	0.97	0.60	0.90	0.88
	3	0.49	37	37	0.63	0.96	0.44	0.84	0.82
γ=.90	100	0.78		1.1					
	40	0.76	1.8	2.8	0.98	0.99	0.96	0.96	0.99
	20	0.73	6.0	7.0	0.94	0.99	0.84	0.91	0.97
	10	0.67	14	15	0.86	0.98	0.71	0.89	0.93
	5	0.53	31	32	0.69	0.97	0.52	0.86	0.86
	3	0.39	50	51	0.50	0.97	0.36	0.72	0.77

Power falloff (column 3) and Efficiency Formula figures are compared to values when m=100. Power falloff figures in column 4 are compared to equivalent FIML model. Falloff figures of "0" in column 4 were very slightly positive (greater power), and were fixed at 0. Power for equivalent FIML analysis was also 0.7839 (for all levels of γ).

simulation as the number of imputations dropped from m=100 to m=3. When γ was small (γ=0.1), the power falloff was not dramatic. For γ=.1, the power falloff was less than 1% with m=40 or 20, but was somewhat larger for m<20 1.4, 1.9, and 3.7% for m=10, 5, and 3, respectively). For γ=0.3 the power falloff was less than 1% for m=40 and m=20, but was 3.4, 7.3, and 13% for m=10, 5, and 3, respectively. In comparison with the corresponding FIML model, the power falloff figures were very slightly lower than the falloff compared with m=100.

On the other hand, for $\gamma \geq 0.5$, the power falloff was noticeable, even with 20 or more imputations. When γ=.5 the power falloff was less than 1% for m=40, but was greater than 1% for m<40 (1.2%, 4.2%, 12.7%, and 24.9% for m=20, 10, 5, and 3, respectively). For γ=.7, the power falloff was just less than 1% for m=40, but was 3.8%, 8.5%, 22%, and 37% for m=20, 10, 5, and 3, respectively. For γ=.9, the power falloff for $m \leq 40$ was greater than 1% (1.8%, 6%, 14%, 31%, and 50% for m=40, 20, 10, 5, and 3, respectively). In comparison with the corresponding FIML model, the power falloff figures were slightly higher than the falloff compared with m=100.

Estimation of γ

We have shown in our simulation that the power falloff was relatively modest when $\gamma \leq .3$. In fact, one might believe, from MI theory, and from our simulations, that when $\gamma \leq .3$, one really can get by with a smaller number of imputations. One problem with this argument, however, is that γ itself is

not reliably estimated unless m is rather large. Table 4 shows the estimates of γ for various levels of γ and m.

One can see in Table 4 that one standard deviation above the mean for true γ=.30 and m=5 is γ=.50. However, the consequences are relatively minor for thinking one's γ is higher than it really is. If one believes erroneously that one's γ=.50, then one simply asks for more imputations, and all is well. However, if one believes erroneously that one's γ=.30 when it is really .50, there could be unacceptable loss of power. Thus, we argue that the most important values of γ in Table 4 are .50 and larger.

As shown in Table 4, when true γ=.50, with m=5, one will estimate γ to be as small as .34 a non-trivial amount of time. When true γ=.50, even with m=10, one will estimate γ to be as small as .40 some of the time. When true γ=.70, with m=5, one will estimate γ to be as small as .50 some of the time.

Discussion

MI vs. FIML

A question is often raised as to which missing data approach is better: MI or FIML. Missing data theorists (e.g., Collins et al. 2001; Schafer and Graham 2002; Graham et al. 2003) have argued that MI and FIML are equivalent. Collins et al. (2001), for example, have argued that the two approaches "... will always yield highly similar results when the input data and models are the same, and the number of imputations, M, is sufficiently large."

The Collins et al. (2001) article focused mainly on the idea that MI and FIML approaches yield similar results when the same variables are taken into account. This issue applies mainly to the idea of including additional variables in the model to "help" with the imputation; Collins et al. referred to these additional variables as "auxiliary" variables. With MI, adding such variables to the missing data model is easy to do. With FIML approaches, however, Collins et al. noted that the researcher must take extra steps to include these auxiliary variables in the model. Graham (2003) suggested models that accomplish these extra steps for FIML-based Structural Equation Modeling (SEM).

The present article also addresses the issue of whether MI and FIML methods are equivalent. Our results show rather clearly that compared to MI with m=100, MI with fewer imputations can lead to an unacceptable power falloff. An important point of this article is that one can avoid this preventable power falloff simply using MI with more imputations.

But it is also important to compare one's power using MI with a certain number of imputations with power that could be achieved using the equivalent FIML procedure. As long as it is reasonable to assume that power based on MI with m=100 is essentially the same as power based on MI with m=∞, then the power falloff figures we show in our tables also apply reasonably to power falloff with respect to the comparable FIML analysis. Indeed, when γ is small, for example, when $\gamma \leq .3$, power based on MI with m=100 is essentially the same as power based on the equivalent FIML analysis.

However, when γ=.5, power based on MI with m=100 is a little lower than power based on the equivalent FIML analysis. For γ=.7 and γ=.9, the differences are even larger. Thus, when one adds the small power falloff for MI based on m=100 (with respect to FIML) to the power falloff for MI with a smaller number of imputations (with respect to MI with m=100), the total power falloff with respect to FIML is slightly larger overall. This overall power falloff with respect to the equivalent FIML analysis was shown in Table 3.

Table 4 Estimates and variability of γ

Population γ															
m	0.90			0.70			0.50			0.30			0.10		
	−1 SD	γ	+1 SD	−1 SD	γ	+1 SD	−1 SD	γ	+1 SD	−1 SD	γ	+1 SD	−1 SD	γ	+1 SD
100	0.89	0.90	0.92	0.67	0.70	0.74	0.46	0.50	0.54	0.27	0.30	0.34	0.08	0.10	0.12
40	0.88	0.90	0.92	0.65	0.70	0.76	0.45	0.51	0.57	0.25	0.31	0.36	0.08	0.10	0.13
20	0.87	0.90	0.94	0.64	0.71	0.78	0.43	0.51	0.60	0.24	0.31	0.38	0.07	0.11	0.14
10	0.85	0.90	0.96	0.61	0.71	0.82	0.40	0.52	0.65	0.21	0.32	0.43	0.06	0.11	0.16
5	0.80	0.90	1.0	0.55	0.71	0.88	0.34	0.53	0.71	0.18	0.34	0.50	0.04	0.12	0.20
3	0.68	0.86	1.0	0.45	0.69	0.93	0.27	0.52	0.78	0.13	0.35	0.57	0.02	0.13	0.25

"−1 SD" means one standard deviation below the mean for γ for that level of γ. "+1 SD" means one standard deviation above the mean for γ for that level of γ.

Recommended Number of Imputations

The simulation results shown in this study are interesting, and have important implications for prevention scientists. Based on these results, we advise users of multiple imputation to ask for many more imputations than has previously thought to be needed. How many imputations are needed depends on γ, to be sure, but also on one's tolerance for the (preventable) power falloff due to choosing m to be too small.

Our recommendations for number of imputations are summarized in Table 5. We begin with the assumption that the tolerance for a preventable power falloff will normally be low. When statistical power matters most, for example, we would require that the preventable power falloff be less than 1%. We also start by comparing our analysis with the corresponding FIML analysis (the rightmost column in Table 5), which is equivalent to an infinite number of imputations. With these assumptions, we recommend that one should use m=20, 20, 40, 100, and >100 for true γ= 0.10, 0.30, 0.50, 0.70, and 0.90, respectively. It could be argued that one should use these conservative recommendations even if a FIML approach is not an option.

On the other hand, there may be situations when one wishes to compare the power falloff with a large number of imputations, say m=100. Also, there may be situations in which one is willing to tolerate a greater power falloff. These situations are captured in the left three columns of Table 5. For example, if one is willing to tolerate a 3% power falloff compared to m=100, then one should use m= 5, 10, 20, 40, and 40 for true γ=0.10, 0.30, 0.50, 0.70, and 0.90, respectively.

In sum, our simulations results show rather clearly that FIML is superior to MI, in terms of power for testing small effect sizes, unless one has a sufficient number imputations. The number of imputations required is substantially greater than previously thought. The number of imputations required for equivalence with FIML procedures is dramatically higher than previously thought when the fraction of missing information (γ) is very high.

Table 5 Imputations needed based on the fraction of missing information (γ), and on tolerance for power falloff

Acceptable Power Falloff				
	Compared to m=100			Compared to FIML
	<5%	< 3%	< 1%	< 1%
0.1	3	5	20	20
0.3	10	20	20	20
0.5	10	20	40	40
0.7	20	40	40	100
0.9	40	40	100	>100

Implications for Large and Small Effect Sizes

The results of this study were based on one effect size (β= 0.0969—a "small" effect size in Cohen's 1977, terms). With larger effect sizes, the power falloff as described in Table 3 would be much smaller. However, selecting the number of imputations in a study is a bit like selecting a sample size. A change in sample size of say, N=500, may have relatively little impact on the power to detect large effects in a study, but it may have a meaningful impact on the power to detect small effects. Similarly, with multiple imputation, the power for testing larger effects may be relatively unaffected by the m chosen for the study. However, smaller effects will be materially affected by the choice of m. Most prevention researchers go into a study with the idea that various effects, large and small, will be tested. If one wants all of one's hypotheses to be tested with good power, then one must pay close attention to power calculations for the smaller effects in the study.

Final Thoughts

In this article, we recommend that researchers using multiple imputation should use many more imputations than has previously been recommended. One might conclude from these recommendations that multiple imputation is no longer a useful tool for dealing with missing data. However, two facts about multiple imputation must be taken into account in deciding upon the usefulness of this tool. First, how much additional computational effort is really required between, say, m=20 imputations and m=100 imputations? In our experience, some analyses do require considerable time, and multiplying that time by 100 would represent a substantial increase in computational effort. On the other hand, in our experience, many analyses (e.g., multiple regression analyses and structural equation models with continuous data) take just seconds to run, sometimes just a fraction of a second. Multiplying this computational time even by 100 represents a trivial increase in overall computational effort. Further, the issue of computational speed very likely will become less important (1) as the computers become more powerful, and (2) as analytic software becomes more efficient.

The second fact relates to the ease with which auxiliary variables (variables highly correlated with the variables of interest, but not part of the model to be tested) may be incorporated into the model. Although it is possible to incorporate any number of auxiliary variables into FIML models (e.g., see Graham 2003; for suggestions regarding SEM-based FIML models), doing so becomes very tedious as the number of auxiliary variables increases. Further, latent class, and other categorical variable models are becoming more common. However, to date, there have

been no published works describing how to incorporate auxiliary variables into these models. Ease of incorporating auxiliary variables into one's model is also likely to become less of an issue over time. Future versions of FIML-based software will very likely include features that allow one to incorporate important auxiliary variables into one's model as easily with FIML as can be done currently with multiple imputation.

Taking these two facts into account, we argue that multiple imputation and FIML procedures will both remain highly useful analytic tools for dealing with missing data. We encourage researchers to make use of both of these important tools.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330–351.

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10*, 80–100.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In: J. A. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 87–114). Volume 2 of *Handbook of Psychology* (I. B. Weiner, Editor-in-Chief). New York: Wiley.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review, 95*, 49–69.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.

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

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571.