# Missing Data Analysis: Making it Work in the Real World

### with John Graham

### John Graham

### **October 20, 2010**

Host [Michael Cleveland](http://methodology.psu.edu/people/mcleveland) interviews [John Graham](http://bbh.hhdev.psu.edu/fac_staff/faculty.asp?id=10), Professor of Biobehavioral Health and Human Development & Family Studies at Penn State, to discuss his recent *Annual Review of Psychology*article "Missing Data Analysis: Making it Work in the Real World." Please contact Dr. Graham at[jgraham@psu.edu](mailto:jgraham@psu.edu) if you would like a PDF copy of the article.   Podcast is available as a 2-part download.

Part 1 of 2

Speaker 1: Methodology Minutes is brought to you by The Methodology Center at Penn State, your source for cutting edge research methodology in the social, behavioral, and health sciences.

Michael: Hello, and welcome to Methodology Minutes. Today I'm interviewing Dr. John Graham. John is Professor of Biobehavioral Health at Penn State University and will speak with us today about missing data analysis.

He recently published an article in the Annual Review of Psychology called "Missing Data Analysis: Making It Work in the Real World," which I think presents an excellent summary of this issue, and even more importantly, provides scientists with some great practical suggestions and strategies for dealing with this very real and important issue.

John, welcome to Methodology Minutes.

John: Thank you.

Michael: Just a little bit first to talk about your background and your biography. You were originally trained as a social psychologist, is that right?

John: Yes, one of my first real jobs, right. Actually, I was still a graduate student in Social Psychology, and I got a job working at an institute that eventually changed its names two or three times, and it eventually became The Institute for Prevention Research at the University of Southern California.

Michael: Okay.

John: It was interesting that many of the people that were there at that time were also social psychologists, so one of the main ideas of that group was to translate social psychology theory into prevention work, preventing drug abuse in adolescent populations. My training, and I think most of the other people that were there at the time, was in experimental social psychology, so a good bit of our work was focused on translating the kind of controls that you can get in an experimental laboratory study, into real-world settings where the controls are not nearly so good. We had to solve all kinds of methodological problems, and one of those main ones was missing data analysis, and I just found that the missing data analysis was one of the ones that I really found interesting and very challenging, and it's the one I've been focusing on for the last 15 or 20 years.

Michael: Hm. Okay. You kind of anticipated my next question, or my question which you've already answered, which was how did you get interested in missing data, so I think I'll just jump right in now and say that I would like to use your review article as the basis for our discussion today.

First, I want to start by telling the story of a colleague of mine who faced one of these real world missing data issues. My colleague received reviews of an article that was submitted for publication in which she used one of the modern missing data methods that we'll talk about later, called F-I-M-L, FIML. She was surprised when one of the reviewers wrote that they were skeptical of these analyses and actually requested to see results when using only data with no missing, what's called listwise or casewise deletion. My colleague responded to this request by noting several advantages of FIML over listwise deletion specifically, and citing several sources including, I think, some of your manuscripts, but felt obligated to still respond to this reviewer's request to report the listwise deletion results. What are your thoughts about this situation? How would you respond in that particular scenario?

John: Well, sledgehammers work really well, but seriously, I think that that idea, fortunately, we're seeing fewer and fewer responses from reviewers that are like that one. They're still not unheard of, and some reviewers still want that kind of evidence. I think the logic of that kind of response comes from a pretty old but well-established idea in evaluation and science, for that matter, the idea of triangulation. The idea is that, if you can get two views of the same thing that we don't understand very well, and if those two views happen to agree with each other, you feel a lot more comfortable about the interpretation that you would make.

Unfortunately, that particular approach, that triangulation approach, even though it's really good in some circumstances, really doesn't apply here. We know, from missing data theory and from just numerous simulations, that the procedures that your colleague was using, Multiple Imputation and FIML, those procedures are known to be better than the old procedures that include casewise deletion.

I think one of the issues that comes up here is that if both of the procedures yield the same results, then you've got confidence. Everybody will be more confident in it, but what happens when they're different, and that's the real key. When they are different, that's where you have to decide, well, what's right? Well, again, from missing data theory and from just numerous simulations, we know that the Multiple Imputation and FIML approaches are going to be better than the complete cases or listwise deletion approaches.

In the 2009 paper, I argued that the MAR methods, or the Multiple Imputation and FIML methods are always at least as good as the old methods, and usually they're better, and often by a lot.

Michael: They're never worse?

John: They're never worse.

Michael: They're never worse, they're at least as good, and most the time, they're better.

John: Right, right.

Michael: Okay, so we're already jumping to the final conclusion, but let's get to that point and talk about the details of that 2009 article.

You divided that paper into four major sections. I would like to at least like to try to cover a little bit of each of those sections today in our broadcast. Those sections, first, are Missing Data Theory. Second, you talk about Practical Issues and Real-World Applications of Missing Data. The third section is devoted to Attrition And MNAR Missingness. Then finally, you wrap it up with a Summary and Conclusions, so let's start with that first section of Missing Data Theory. You spend quite a bit of time in this section, explaining what you call Causes or Mechanisms of Missingness and note the looseness that surrounds these terms. Can you explain what you mean a little bit there?

John: Yeah. The two kinds of missingness that we have, the big division is between missing at random, which is often referred to by its acronym, MAR, and missing not at random, or MNAR. Really, the other kind of missingness, missing completely at random, is a special case of MAR, so we can kind of leave that alone for a minute.

Michael: Okay.

John: The biggest issue here, is that there really is no such thing as pure MAR, purely missing at random. There's no such thing as purely missing not at random. I would argue, and do, in fact, that there is a continuum, and the reality is always somewhere in the middle. My quibble here is that statisticians, who developed all of these ideas, they know that this is a continuum, and yet, many people who talk about missing data act as though it's one or the other, and sometimes it's a bit of looseness with regard to just the English language. A lot of times a question might come up about someone saying, "Well, what if the cause of missingness is MNAR?" Well, that implies that it is possible, and that the cause of missingness would be missing not at random, and I would argue that that's just never the case.

Anyway, I'll talk more about that a little bit later, but one of the big things that I try to do in the 2009 article, but I'm also just trying to have people think about is that, whether it's MAR or MNAR isn't even something we should be thinking about. We should focus, instead, on the consequences of those things, and think about, whatever it is, what happens to the bias? Are the results biased, and by how much, and are our results tainted because of the missingness, and can we trust the results? Those are the kinds of questions that we should be asking, rather than whether it's MAR or not.

Michael: A strict yes or no, or dichotomous category of the type of missingness then.

John: Right.

Michael: Okay. Can we talk a little bit now what you describe as the older or more traditional ways of dealing with missing data as a good place to start and that we can build from?

John: Sure. Probably among the old methods, complete cases analysis or listwise deletion is really the most common. This is the case where the subject has, let's say there are ten variables that you,... A small questionnaire has only about ten variables in it, and if someone has nine of the variables but it's missing one, you'd drop that particular subject from the analysis then. Only the cases that had complete data for all the variables involved in the analysis would be kept.

One of the things I like to say sometimes, is that this method is actually not so bad if you're not losing very many cases. In some of my earlier work, I went on record as saying that if you're only losing about five percent of your cases due to missingness, you probably weren't going to be very far off with complete cases analysis, because how much bias can there be with only five percent missing, and how much loss of power can there be if you have only five percent missing, so it didn't seem like you were doing something completely horrible.

Now I've kind of taken the other position. I want people to use the better procedures, and I think they're always going to be better, and I like to nudge people towards using the better procedures, even when there's very little missingness. That's one of the old methods.

Another old method is pairwise deletion. This is a procedure that's used when you're working with a correlation matrix, and all of the estimates, the correlations and the variances, will all be estimated based on the people who have data for that particular parameter.

Michael: Okay.

John: What ends up happening is if you would have a different subset of cases, you would have data for one correlation versus another correlation, and that can be a bit of a problem, that idea that you're talking about different subsets of people. The biggest issue is that with pairwise deletions, there's no way to estimate standard errors for the correlations, and that can be a big problem.

Then finally, I almost hesitate to say this, the other old method for taking care of missing data was known as means substitution. With means substitution, of anybody who had data, you'd get a mean on a particular variable, and then if someone was missing on that particular variable, you would just substitute that mean. Let me just say that that's a horrible thing to do. Don't ever do that. Don't do it even if you've got one value missing. Don't do it.

I have this thing I say, and I'm sorry if it sounds a little mean, but if you want to pretend that you have no missing data, I have a better way to do that than this, and I'll talk about it Later.

Michael: Okay. Will you tell us, eventually, why that's such a bad idea?

John: Oh, the biggest problem is that the variances and the correlations are both wrong, are both biased, substantially biased, and there's no real way to estimate standard errors, so you have biased estimates and no standard errors, so it's a very bad situation, it's the worst of all situations.

Michael: You're in the worst case scenario. Okay.

John: Listwise deletion can be very useful under certain circumstances. I even still use pairwise deletion under very specific circumstances. I would never publish off of it, but sometimes I use it to get a very quick idea of what's going on in the data, and sometimes it can be useful. Means substitution is never.

Michael: Okay. Yeah. I think you mentioned in the article that you used the pairwise sometimes, if you're doing exploratory factor analysis and you just want to get a sense of what the structure may be.

John: Right. Yes, if it's a very large data set, sometimes the pairwise deletion is the only way to approach it early in the process. I also use it for troubleshooting sometimes, when you're having problems with a procedure, like with Multiple Imputations, for example, if you're having trouble with it, it won't converge, something won't happen, I often will use pairwise deletion factor analysis to get a sense of what's going on with the data.

Michael: Okay. Now that we've covered these older methods that you obviously do not recommend, and some more strongly than others, we should probably talk about the modern missing data methods that you do endorse. Those would include several different strategies. Can you talk a little bit about what those are?

John: Sure. The main approaches that have been, I'm going to just touch the surface here, but the EM algorithm is something that's been used for a long time with respect to missing data, Multiple Imputation, and FIML methods, the Full Information Maximum Likelihood procedures.

Let me talk a little bit about, ... All of these procedures give you something really valuable, which is the ability to make use of partial data. This idea that somebody has data for nine of the ten variables and you have to throw away those nine of ten variables, one variable missing and you have to throw it all away, that just seems too bad. There's a lot of information in that person's data, and you're still throwing it out. All of these procedures help with that. All of these procedures give you ways of making use of the partial data, and that's really a key thing.

I think that the thing we're trying to get with any analysis procedure is unbiased estimates, and I think that all of these, EM, Multiple Imputation, and FIML methods all give us unbiased estimates. In other words, the estimate we get, like regression coefficient, or correlation, or a mean of something, they're very close to the population value, and that's what we're trying to do.

The other thing that we're trying to get from all of these procedures, or from any procedure, is some estimate of the variability around that estimate, like standard errors. You want to know, is that estimate really close, or are you pretty sure that it's in that area, or are you not so sure that it's in that area? Especially in the missing data case, the biggest thing we're trying to do is to make sure that we model and understand the uncertainty due to the missing data, and we need to make sure that that's part of the estimate, and that uncertainty needs to be modeled.

Michael: Okay. You've described this as a situation where you're missing, let's say, one variable for one case, that case that had nine of the ten, and you're able to, I guess, use that partial data in your analysis, or eventually. Doesn't that mean that you're making up data or giving yourself some kind of edge? Where do these imputed numbers come from that you're talking about now?

John: Yeah, that's one of the questions that gets asked a lot, and again, that's one of those things that I'm happy to say, the question comes up less and less these days. I think that's because our message is starting to get through, but the idea that we're somehow helping ourselves by filling in the blanks so that we can make use of the partial data, we are not helping ourselves. When you do these kinds of analyses, you're not getting something for nothing. All you're doing is just allowing yourself to make use of the partial data, and you basically get everything you deserve. Every time I say that. I think it has some other meaning, but you get everything that's due you, in the statistical sense.

Michael: Okay.

John: One other thing that comes up that's related to this has to do with, people sometimes think that that value that you plug in, that you impute, means that that's what the person would have said had they given you the data, ...

Michael: Right.

John: ... and it's really important to realize that that's not what you're doing here. You're not trying to figure out what the person would have said. What you're trying to do is just organize the data in a way that allows you to make use of the partial data, and allows you to get estimates for the whole population, or for your whole sample, that are unbiased. That's really important, so you fill the blanks in, in a way that allows you to get an unbiased estimate of what the correlation is involving that variable.

Michael: Okay, sure.

John: That's kind of a key there.

Michael: Right. Okay. Going back to these three specific, modern missing data strategies of EM algorithm, MI, and FIML, what kind of comparisons can you make across those three? Are there situations that call for one versus another, or what kind of differences would you say describe those three modern methods?

John: First of all, I think they all have their uses. I really like EM. I think maybe I like it most because, when I was first learning about missing data, that was the first thing I learned. It was such a hard task to program an EM algorithm and come up with a stand-alone EM program. The fact that I worked it out with the help of my colleague, Scott Hofer, it just felt so good, and I've just made a lot of use of EM, so I really feel very fond of EM, as it were. A lot of times, the results of EM algorithm are very good in the sense that you get very good estimates, they're unbiased, they're Maximum Likelihood estimates, or at least the correlations and co-variances, and variances that you get from EM are maximum likelihood, and that's very useful. The big problem with EM is that you don't get standard errors. There's no mechanism for getting standard errors from those estimates, or from the EM approach itself. Still, there are lots of times when you don't really need standard errors, you don't really need hypothesis testing, so it turns out that EM and, ... I'll say that EM-related methods, because it's possible to impute a data set from EM parameters that has a very good property. It has most of the same properties that the EM matrix has itself, but because it's a data set you can do, sometimes it's easier to do certain things with it.

One of the nice things about EM is that you can estimate, like coefficient alpha, if you're trying to figure out what's going on with your scales and estimate your data quality, you can do that with EM. It's a very good way to do that. Similarly, exploratory factor analysis is very good with EM, because you're not usually doing hypothesis testing with the factor loading. You're just trying to get the parameter estimates, and those are very good, so it's a really good way and useful way to get those estimates when you have missing data.

On the other hand, most of us usually are trying to do some hypothesis testing. We want to know if a program worked, or we want to know if two variables are significantly related to one another.

Michael: That's the ultimate question.

John: Right, and that's where EM can help you, and that's where you would switch to Multiple Imputation or FIML. I'll say this again, that FIML is Full Information Maximal Likelihood, but it's just gotten to be so much of a mouthful that I usually just refer to it by the acronym.

Michael: Actually, can you back up a second and maybe just tell the listeners what EM stands for? That might help clarify, maybe, what you just talked about in terms of why this doesn't provide standard errors, and things like that.

John: Well, I think what I'm going to say first won't actually help that, but EM is Expectation Maximization, and that doesn't really help anything either, but I'll give you a very brief rundown of what EM does. It's another thing I like about EM, the fact that I know, because I worked out the program for it. EM, in the expectation, the E step, it's an iterative procedure, and in the E step, you read in the data, and you calculate the fundamentals of correlation and regression. Namely, you calculate the sums, sums of squares, and sums of cross products, so it's just the building blocks of correlation, regression, and variances, and those kinds of things. Then, wherever you have data, you read in the data, of course, and wherever you are missing data, you give your best guess, you estimate the best guess, and that's usually a regression-based best guess, so it's a single imputation of that variable being predicted by all the other variables.

When you're all done with that, you have the sums, sums of squares, and sums of cross products in the sample size, you can now estimate, you can actually calculate the variances and the co-variances. Based on the variances and the co-variances, you can calculate the regression coefficients that can be used for further imputation in the next step, so the calculating of the co-variance matrix, that is the M step for the first iteration, so now you use that new co-variance matrix to create the B weights and the regression coefficients for imputing the data in the next iteration of EM, the next E step. Then you get new estimates of the variances and co-variances, and then you go back and forth like that until the estimates of the variances and the co-variances stop changing. Once they've stopped changing, you say that EM has converged, and that convergence is at a maximum likelihood solution.

Michael: Okay. Thank you for that explanation. I think that will lead to what you were going to say about MI next, right?

John: Well, I think that the key thing here is that with the Multiple Imputation, the idea is that you're going to be working with data that the outcome of Multiple Imputation data. You're going to create a data set that has no missing values in it, and Multiple Imputation is when you create that data set over and over and over again. There are a couple of reasons why, when you do single imputation, that the variance is too small and you want to restore some of that lost variance, and Multiple Imputation does that in a couple of ways. I'm not going to get into the details of that here, but ...

Michael: Okay.

John: ... you do restore that lost variability in that the result is estimates of variances and co-variances and regression coefficients and pretty much anything that you want that are all pretty much unbiased, and you get a sense of the uncertainty due to missing data because of the way it's all put together. You can actually get an estimate of the uncertainty due to sampling, which is the usual kind of uncertainty, and then you get an estimate of the uncertainty that's due to missing data, and put those two things together and build the overall standard error.

Michael: Mm-hm. Okay.

John: The FIML approach takes a whole different procedure here. It's more of a one-stop shopping. You estimate the parameters of interest, like regression coefficients, at the same time that you are taking care of the missing data, so it's all one thing. The algorithms themselves have all been rebuilt, so the result of the FIML approach is that you do get the parameter estimates, reasonable standard errors, all in a single step, all in a single analysis. Most of the implementations of FIML have come from structural equation modeling. That's really the most popular one. There's a couple of other approaches that do FIML as well, but structural equation modeling is clearly the dominant implementation of the FIML procedures.

Michael: Mm-hm. I think it leads to the next section that you discussed in the article, which is regarding the practicality of these methods. Can most scientists easily incorporate these procedures into their everyday work, and are there software packages that can easily do these methods for us?

John: Absolutely. This actually I see as my forte in this area.

Michael: Great.

John: I'm not someone who develops the procedures. I didn't develop Multiple Imputation. Don Rubin, I guess, is credited with Multiple Imputation, and there's numerous really great statisticians out there who have contributed to the development of Multiple Imputation and the Maximum Likelihood procedures, and there are a variety of them. My area is taking that information and trying to bring it to the masses. It allows you to do things that make it easier for people to use.

Speaker 2: Sure.

John: First of all, the FIML methods have really become popular in structural equation modeling packages. Pretty much every structural equation modeling package that's used these days has a FIML feature that allows the personal to handle the missing data all at once.

Multiple Imputation is now really taking off. Joe Schafer's NORM program was developed after his 1997 book on Multiple Imputation, and his software, I think, has really been the beginning of the real blossoming of Multiple Imputation as a practical tool.

I think what happened after Joe wrote his book in 1997, and after it was clear that his program was working pretty well, SAS, the Statistical Analysis System group, adopted his algorithms, and they came up with something called PROC MI, which is just a really amazing program. When it first came out, it wasn't so great, but I think this third version of it now, it's really, really a mature, normal model-based imputation program, and it's very, very useful.

In fact, even SPSS, in Versions 17 and 18, has developed a Multiple Imputation feature, and I have to say that their first try at this, in Versions 17 and 18, are not quite there. The imputation part is still lacking in some really important ways, but there's two things that the Multiple Imputation thing does, and the first one is to do the imputation. The second one is to allow the user to do the analysis and to combine the results in a very automated way, and that second part works really well in Versions 17 and 18 of SPSS, so even though the imputation part still needs some work, but the back end of it, the second part, is working really great, and that's, I think, something that's really important.

One of the things that I've been really focusing on in my own work is to try to develop ways of using Joe Schafer's NORM program to impute the data, and then to automate the process of making use of those imputed data sets with a variety of other programs, including SPSS. I've finally developed something that is starting to be very useful and very automated with respect to Versions 17 and 18 in SPSS. Even with some help from people at the Methodology Center, I'm even getting a Windows-based interface so that the user just answers a couple of very quick questions on a Windows window, and it just prepares the data perfectly for SPSS, and once it's prepared, everything is just business as usual.

Michael: Yeah, I think that sounds like a great advance, because I know I have used NORM in the past, and that was kind of a rate-limiting step, that after you'd run NORM, you had data, but it wasn't ready to analyze into the other package, so that might have put off some users and maybe pushed them more towards the PROC MI and SAS.

John: Well, PROC MI and SAS is really a good product. I have no qualms at all about telling people, "If you're a SAS user, you don't have to go any farther than PROC MI, because you're getting everything you need from that, but if you're an SPSS user," I said, "the procedure isn't quite ready yet, so you really do need to use another procedure."

I also have other utilities for other programs, and many programs have their own utilities these days, but I have a utility for HLM multilevel program, I have utilities for all of the structural equation modeling programs for doing Multiple Imputation with them, with EQS, with LISREL, and Mplus, and all of those can be useful as well, but I think the one that really makes the most sense is the one for SPSS right now.

Michael: Right. That makes sense.