# 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 2 of 2

Michael: Another issue that you talk about in this section of practical issues regarding missing data is the use of auxiliary variables. Why would you say that's important?

John Graham: This is an interesting issue. This is one that I just find so interesting. The idea with auxiliary variables is that ... Well, you start off with a situation where you have some partial data like, for example, let's just say you have a program variable and an outcome ... Let's just say cigarette smoking. Let's say you have a 1,000 cases who have data for the program variable but due to attrition or whatever, you only have follow-up data on the smoking measure for 500. You've got 500 missing cases. That means you've got missing information on them. The parameter estimate that you're trying to get ... The regression weight, the B weight from ... Program predicting of smoking, you've got missing information on 500 cases. Half the information is missing onto that.

You can kind of imagine ... Suppose you were to say "Okay, well, I didn't get those 500 people in the main measure. What if I follow them up and I track them down and I get the information. I get the data on them." The data you would collect on them would be pretty much the same as if they had been there in the first place. The idea is that now you have this other information that is used in place of the information you didn't have ... The lost information from the missing data. We can see from that situation that you've restored a lot of the lost information.

One of the things that makes that work is that the correlation between the data pick-up after the fact and the original depended variable measure of cigarette smoking, that that correlation is really high. Whenever you have a variable like that that's correlated really highly with the variable of interest that has missing data, you can restore some of that lost information.

This new variable, this new variable we collected ... That's an auxiliary variable.

Michael: Oh, okay.

John Graham: The nice thing about this is those variables can come from a lot of different places. What I just described was a situation we like to be able to do a lot, but we can't really do it. A lot of times we could kind of do it the other way around. Sometimes the variable will already exist in our data set that we can use as a good auxiliary variable.

Good example of this is so it was we have a program variable and say in 7th grade, we delivered a program and then in 9th grade, we were going to measure the final outcome of cigarette smoking but 50% of the people were missing.

Michael: Right.

John Graham: Well, what if we went back to 8th grade and went back to 8th grade smoking, we'd probably already have that variable because it's a longitudinal study ... Channel study where we measure the same people over time.

Michael: Sure.

John Graham: What we'd have is 8th grade measure and maybe the most of the people have data for 8th grade smoking. We'd look at the correlation between them. We'd say "Whoa, correlation." It's not one but it's probably pretty high. It's probably like .7, maybe even as high as .75. What that means is that by incorporating that 8th grade measure of cigarette smoking into our analysis, this auxiliary variable, we can actually restore some of the lost power to our analysis.

Michael: Sure.

John Graham: I think that's the key. Restoring lost power and actually lowering any kind of estimation bias that might come from the missing data on the 9th grade smoking variable.

Michael: I see.

John Graham: Restore lost power and reduce estimation bias. Auxiliary variables are our friends.

Michael: Okay, good. We should include those. Another kind of real world, practical issue regarding these strategies that you discuss in the article has to do with including interaction terms. That can, maybe, seem daunting to someone who's trying to use these methods. When you do an analysis, let's say, and you use imputed data and you initially decide, "Oh, I think I want to interact gender with this particular independent variable," what are the implications for that in terms of these, particularly MI strategies?

John Graham: Yeah, this is a big issue. One of the basic rules of multiple indication is that any variable that's going to be in the analysis, should be in the imputation model. That means that if you have a variable, an interaction that you would like to include in your analysis, you better include that interaction in your imputation model. As you say, that can be kind of daunting, especially if you've already done the imputation and you realize, "Oh, gosh, I want to add this interaction with gender." You have to go back and re-impute and include that interaction in the model.

A couple of things you can say is that you would need to ... Well, let me first say why that's important.

Michael: Okay.

John Graham: One of the things that happens is that any variable that's omitted from the imputation model ... So if a variable is omitted from the imputation model, then the imputation is done under the model that the correlation is zero between the omitted variable and all the variables that are in the model.

Michael: Okay.

John Graham: That often is a very bad thing especially if you're trying to find an interaction ... Big time interaction in the imputation model and then you decide later you want to do it, you're suppressing the results of that towards zero. You're making it even less likely to find it. That can be hard to do and hard to anticipate. You'd really have to sort of anticipate every product term and throw them into the imputation model and also that has a lot of variables. Fortunately there's one approach to this that ... Especially for certain classes of variables that can be very helpful and relatively manageable. You can just take the gender example first.

One of the things you can do is to impute separately within the two genders. If you did separately within males and females, then what that does is that it preserves the mean difference between and all the correlation differences within males and females. You can have differences in correlations and that is the definition of an interaction.

Michael: Okay, sure.

John Graham: Different correlations in the two groups. Because you're allowing those correlations to be different, you're not pushing them to be the same ... Then that means that you essentially have included all possible interactions with gender in your imputation model. Making one split like that, males within males within females, is usually possible without too much trouble.

Michael: Sure.

John Graham: It's when you get too many ... Like if you wanted to add program. Program is another one ... Program within program within control. If you want to do the program males, program females, control males, control females ... Now you're starting to get a little bit ... A 3-way interaction

Michael: Very complicated.

John Graham: It's harder and the bigger it gets, it gets hard very fast.

Michael Oh, okay.

John Graham: You can certainly do regular interactions pretty easily this way.

Michael: Okay, good. I also wanted to talk a little bit about ... It seems like your discussion so far is focused on normal, continuous data. How well do these strategies work if you have other types of data which, for many researchers particularly in substance use and prevention, will probably be ... They may be categorical or definitely eschewed and maybe not normal? What kind of advice do you have for those users?

John Graham: I think the bottom line is that normal model multiple imputation works pretty well with everything. It's always best if you have data that don't conform to the normal model. It's going to be best if you can develop an imputation model that really is tailored to that kind of data. Categorical data is a really good example. There are models. One of Joe Schafer's models is called the CAT program. It is for categorical data.

The problem we face a lot times is that the other models are not so well known and they're not as available as the normal model multiple imputation. What you end up having to do is to use what you have available. Fortunately it turns out that normal model MI works pretty well with everything. It works pretty well with categorical data. Two level categorical data, for example, is no problem at all. If you have a very light gender ... Just got zero in one, you just put that into the model as if it were continuous and the model works just fine. It's all based on correlation. Any variable for which a correlation is meaningful, you can just include it in the model.

Michael: Sure.

John Graham: If you have a categorical variable that has more than two levels, you have to dummy code the categorical. If you've got four categories, you would have to include three dummy variables. The three dummy variables would then be included in the imputation model.

Michael: Right.

John Graham: You can always reconstruct it back to its original form after imputation. With the two level variables, the two level categorical variables, you can impute those if there's missing values. One of the things that happens is that ... You can actually do the imputation with more than two levels as well. One of the things that happens and it's just a normal thing ... That the imputed value almost never falls right on the zero or one category. It's usually somewhere else. Sometimes you can just leave them like that. One of the things I often say is that if gender, for example, is just going to be used in the model as a co-variant, it doesn't hurt just to leave it in that form as the co-variant.

On the other hand, if you're going to use it as something where categorical ... It means to be categorical, you're going to group the people according to gender event, that would require them having zero-one categories. That's when you would just round to the nearest legal value.

Michael: Oh, sure. Okay.

John Graham: There are similar things that can be done with categorical variables that have not more than two levels.

Michael: Okay. What about data that are nested or at a higher article in terms of clustering?

John Graham: Right. That's a really interesting issue. Even though my expertise in multi-level models has always been a little bit shaky, I've always found the imputation of it, of multi-level data, to be very interesting. Again, it's best if you have a model that's ... An imputation model that's designed specifically for multi-level data. Again, that's not particularly available so we make do with what we can. It turns out a normal model MI works really well with certain classes of clustered data.

For example, if we had a situation with students, children in schools ... Children within schools ... We have, a lot of times they have a prevention week due assignment, the random assignment to conditions ... We'll do that at the school level rather than at the kid level.

Michael: Right.

John Graham: Anyway, we go to analyze the data and what ends up happening is that the kids within a school are more like each other than they are like kids outside the school. That's the sort of structure that we want to be able to take into account. You can take care of some of that information by simply dummy coding the school membership. There aren't too many schools

Michael: Oh, okay.

John Graham: You could dummy school membership. That dummy coding, what it does is that it allows the means to be different in the school.

Michael: Sure.

John Graham: You're imputing under the model that allows the means within the schools to be different. You've not constrained the means all to be the same.

Michael: Okay.

John Graham: This is equivalent to what's known as a random intercepts model in multi-level analysis. That approach works really well provided you don't have too many schools. I've handled as many as 35 schools or communities. It seems to handle it really well. You've got to bump that number when you're starting to reach the limit of what most of the multiple imputation programs can handle and then it gets to be more of a problem.

Another issue that comes up with multi-level models though that is not so straightforward is the idea of random intercepts and random slopes models. That's the case where both the means and the covert answers are going to be different in and across the groups. That makes it a lot hard for a multiple imputation to handle because it's really difficult to impute in a way that allows you to allow different correlations within the different clusters.

It is possible to impute with ... One day we're at 12 schools. I could potentially do the imputation separately with the each of the 12 schools and that would preserve the different correlations within the 12 schools.

Michael: Like you talked earlier with gender being.

John Graham: Yes, yes. Exactly the same concept but when you are stretching it out to the number of schools, the sample size within the schools gets to be small.

Michael: Right.

John Graham: So small that the number of variables you can really handle that way gets smaller. It limits what you can do. If you've got too many of those categories ... Like if you've got a 100 schools or something, then it gets really difficult to compute separately a 100 different times and pull the information back together.

Fortunately the biggest use of multi-level models is to correct for the differences in means in the clusters, in the schools, let's say. I think that that means that the most common implementation is the random intercepts model. That really helps.

Michael: Yeah.

John Graham: Because that's the most common implementation. That means the multiple imputation solution looks very well.

Michael: Okay. Another common situation that you devote actually an entire section of your book to is regarding attrition with longitudinal studies. Can you tell us a little bit about what you suggest to the listeners today who are facing this very common problem of maybe 50% of their sample is gone by way 4 or way 5. Are there a couple suggestion you could offer for them? I know you've devoted an entire section that would be hard to cover in a few minutes but what's the bottom line in that case?

John Graham: This is really interesting. If I see my career going in any direction over the next few years, it's going to be this ... The study of attrition and issues surrounding it.

Michael: Okay.

John Graham: This is a really interesting topic. One of the things I would start with is something I said a little while ago when we're talking about missing at random and missing not at random.

Michael: Yes.

John Graham: Think of that as being a continuum. Attrition is always sort of in this ballpark. People have always feared that the attrition that they have, especially when it's as high as you're saying ... 50% ... A lot of people, half you're people are gone at the end of the study. If the missing-ness is not at random, oh, you really worry about how that's going to impact your results.

Michael: Sure.

John Graham: I want to start with this idea that forget about the idea that the cause of missing-ness is MMAR. The cause of missing-ness is to MAR. Remember that it's somewhere in the middle.

Michael: Okay.

John Graham: Instead of doing that, what I want to do is just focus on this idea that we're going to focus on the bias. We're going to say, "Okay, how much bias are we getting due to the missing data, the attrition that we're seeing in this context?" We can usually figure that out. We can get a sense of how much bias there would be and we can do simulations that sort of model the situation that we are seeing in our data.

One of the things I tried to do in the 2009 paper was to develop the beginnings of a taxonomy for attrition. I really think that's what we need to be doing here. I came up with 8 categories, 8 ways of thinking, 8 contexts, you might say, of attrition. I think we can spend some time working on those 8 and try to say, "Well now if we had this situation under these circumstances, what would the effect on our statistical conclusions be?"

Michael: Okay.

John Graham: One really important study was done recently when I called on Schafer and Khan, really worked very hard on one of those 8 categories. They showed some of what I thought, really interesting things. They showed very small amounts of attrition bias even with 50% attrition. That was the thing that really opened my eyes. I'm starting to say, "Whoa! Under those circumstances, that's not bad at all. That looks pretty darn good for our statistical conclusions." That was really what got me started down this road of realizing there's not just one but there's like 8 different ways that you're going to have missing data due to attrition. If we can model that in simulations, we could really figure out what the problem is.

The bottom line is we need to know when the attrition is really there ... The bias is really there and it's really going to affect our conclusions or when we can refer to the bias as being or having a tolerably small effect on our statistical conclusions. I think we're going to find that that tolerably small circumstance really comes up a lot more often than what we used to think.

Michael: Okay.

John Graham: I think one of the things I do in the 2009 paper is that I really, I admonish people who are doing simulations to be very careful about choosing the parameters of their simulation so that they really reflect reality. As much as ... I've gone through this myself, I know I've chosen some really weird parameters for a simulation so I know what can happen. Sometimes the parameters that have been chosen in simulations that have been published are just not anywhere near close to reality. They give a sense of the problems with attrition that are kind of overstated. What I admonish people to do is to really pick very realistic parameter values so that the results will really be helpful.

Michael: Finally, I noticed in your 2009 article that you also talk about what I may consider kind of a crazy idea which is actually incorporating the same data into the design of a study. That seems a little crazy because we just spent the last several minutes or more talking about how to deal with missing data and now you're suggesting that maybe from the very beginning, a researcher can start out with missing data. What's the story behind that?

John Graham: Boy, the number of times I've heard that statement. Why would you want to have a cause to see data? Why would you want to design it in?

Michael: Exactly.

John Graham: Topic of planned missing data design is something that I really like the idea of this. This first came up in 1982. I had only been at this research institute for a year or so. We just got the first large scale, multiple drug use prevention study funded by the national institute on drug abuse. We were going to be in this Los Angeles unified school district ... This huge study, really large. We had four people working on the study, four PIs. I'm counting myself as the fourth one. I wasn't a PI at the beginning of that study. There were at least four of us writing questions on the questionnaire.

Michael: Okay.

John Graham: Our questionnaire was like 300 items long.

Michael: That's a big, old survey.

John Graham: Way ... yeah, right. Very typical and especially when you have multiple people investigators.

Michael: Right.

John Graham: What are you going to throw away?

Michael: Everybody wants their questions.

John Graham: Yeah, right.

Michael: They're dear to their hearts.

John Graham: Right. We just couldn't get the questionnaire down to an amount small enough that the kids would respond to it. I came up with this hair-brained scheme that we could actually have three forms of the questionnaire and that a third of the questions ... We had the big block that we were going to do this with. One of the block, everybody got those questions on say, substance use ... Recent substance use. All the other questions in the questionnaire, 1/3 of the people didn't get one set. Another 1/3 didn't get another set and another 1/3 didn't get another set. By doing it that way, we were actually able to estimate correlations for everybody on every ... We were actually able to estimate correlations on every pair of variables in the data set.

The thing that made that really good was that we were able to collect data on 1/3 more questions than we could have done if we just used a one form design. Of course, that was a very tough sell. I remember the people that I was working with at the time said, "Oh we can't do that. We can't do that." After I came up with ... After I suggested three or four more times, they said, "No, it doesn't sound like such a bad idea." I was thinking about being kind of your backed-against-the-wall situation that you desperately want to ask this number of questions but you don't have time. You're subject to either won't or can't answer that many questions.

Michael: Sure.

John Graham: So what are you going to do? One option is just to ask fewer questions and that's it. The option is to ask fewer questions the way ... In my three form design, you ask fewer questions but in using this design would mean you still get to ask the same the same questions.

Michael: Right.

John Graham: You still get to collect data on all those questions.

Michael: Right.

John Graham: The whole point is to get more bang for your buck. Collect data on more questions given a limited amount of resources. I have to be very honest. When I first started out doing this, I had no clue. This is 1982. I had no clue how it was going to analyze the data. I just thought, well, this ought to work. About 10 to 12 years later, I finally came up with the method for analyzing it.

Michael: Okay.

John Graham: Proved to be very good. Perseverance.

Michael: Very good. I think that's probably a nice place to end our conversation today with kind of, maybe just a summary statement. Can you give us the bottom line now of what you think about missing data and what researchers in the real world face and what they can do?

John Graham: I think people ought to use the missing data procedures that we've developed. Maximum likelihood procedures and the multiple imputation procedures and even occasionally the EN algorithm and related procedures. We should use those. Those are available. Everybody should make use of those. The fact that SAS and now even SBSS has those things and all the structural equation modeling programs have a missing data feature built in, it's just making it easier and easier to use. I want that to be our basis. Everybody gets up that level.

Michael: Yes.

John Graham: And then we can get better from there. It's a really good place to be and I'd like to see everybody doing it.

Michael: Kind of a benchmark for everything.

John Graham: Yes.

Michael: It should at least start there. Very good. Finally, what's next on your plate? Do you have any planned publications that are coming out regarding these topics?

John Graham: My book. I've been working on a book for a little over a year. It's actually called "Missing Data, Analysis and Design." It's contracted with Springer to publish the book. Right now I'm about half finished. I have seven chapters written and approved by the publisher. Figure I'm going to have about fourteen chapters. I plan to work very hard over the next three months and I'm hopeful that the book will be finished this year.

Michael: Great.

John Graham: In the 2010s.

Michael: Wow.

John Graham: I'm looking forward to it a lot. I like how it's turning out and I'm very hopeful that it will be interesting.

Michael: Sounds good. I know I'm looking forward to that. I'm sure that everyone out there will be looking forward to that as well. I'd like to thank you today, Dr. John Graham, from Penn State University, Bio-Behavioral Health, for joining me today. It's been a very good and interesting discussion.

I would like to note in conclusion that listeners to today's podcast can go to our website methodology.psu.edu and find a link to this podcast as well as a list of resources that we will provide for the article that you mentioned today and other sources that our listeners can access.

John Graham: Thank you, John, for enjoying me today.

Michael: Thanks, Michael.

Speaker 3: You have been listening to Methodology Minutes, brought to you by the Methodology Center at Penn State, your source for cutting edge research methodology and the social behavioral and health sciences. This podcast is available on iTunes and at methodology.psu.edu.