Statistica Sinica Preprint No: SS-2016-0474	
Title	Contribution to the discussion of "Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?," by Xianchao Xie and Xiao-Li Meng
Manuscript ID	SS-2016-0474
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0474
Complete List of Authors	David Draper
Corresponding Author	David Draper
E-mail	draper@ucsc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

Statistica Sinica

DISCUSSION

David Draper

University of California, Santa Cruz

This interesting and important paper encourages all of us to expand our standard horizons and consider what Xie and Meng (hereafter XM) call *multi-phase inference*, in which

(a) different teams of analysts (or possibly even the same analysts at different points in time) may be involved in different phases of an analysis, viewed comprehensively from data collection and {data wrangling and curation} to data analysis (possibly consisting of multiple phases itself) and interpretation, but

(b) the statistical models used in some or all of the phases that involve modeling may be based on incompatible assumptions.

I can reinforce the need for multi-phase thinking in contemporary statistical work by relating some of my recent experience in data science at two large eCommerce companies, denoted (for reasons of confidentiality) by X and Y:

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

- In both companies, the end-product of much analysis and decision-making is a web site that can be visited by people wishing to buy or sell various items. This site is supported by a large amount of experimentation and modeling aimed at improving the user experience. Each company has between 10 and 100 groups/teams, working with various degrees of independence from each other, all tinkering with fundamental aspects of how the web site functions (an example at company X is a recommender system to help users either sharpen or broaden their searches for products similar to the one they're looking at now). It's frequently the case that the analytic output of one group forms the input to another group, and it's often true that there is sufficiently little communication between groups that the team receiving an analysis has little understanding of how it was arrived at. It may seem hard to believe that successful companies permit this level of inefficiency of communication and lack of multi-phase thinking, but they do.

- A specific example of failure to adopt a whole-systems perspective at company Y is as follows. There is a team that makes decisions on behalf of the entire company about how the data stream, generated by users of the web site, at its most granular level — time-stamped

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

data about where in the tree of company Y's web pages the user left-clicked his or her mouse, and even spatio-temporal data tracking the location of the mouse arrow to the millisecond — is summarized for analysis by other teams in the company. I discovered that this data summary team had made a statistically unfortunate decision, namely that data that kept track of demand for a particular item in a given time period recorded a 0 for two completely different reasons: a 0 would be entered into the data base that the rest of the company used either if no items were bought or if the item in question was not yet in the catalog of items offered to the users (!). When I inquired about what would be involved in fixing this self-inflicted problem, I was told that it would be politically unwise to pursue a solution, because the data-summary team was under a different Vice President in the corporate hierarchy than I was (!).

Alex Terenin and I have recently been thinking about a framework that includes XM's multiple imputation instance of multi-phase inference as a special case: viewing the output of one team's Bayesian analysis sequentially as the input to the next team can be referred to as *Bayesian model composition*, in the functional-analytic sense that team 1 operates on the available data D, yielding $f_1(D)$, which is then operated upon by team 2,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

yielding $f_2(f_1(D))$, and so on. One question that immediately arises from this perspective is: How can team i craft its f_i in such a way that no important information is lost in the sequential analysis (when compared, for instance, with an ideal all-encompassing Bayesian analysis by a single meta-team)? In extremely simple situations we know that the usual "yesterday's posterior is today's prior" sequential use of Bayes's Theorem accomplishes this goal, but in complex settings it's not at all obvious how to build no-information-loss operators f_i. XM's work can be seen as a detailed attempt to wrestle with this question, in the context of trying to cope optimally with missing data.

- In Section 1.1 XM point out that "... the key issue is that during the journey from God's data to the analyst's data, a set of assumptions have been introduced deliberately or accidentally." I've recently run into a somewhat nonstandard example of this in the teaching of introductory statistics to undergraduates. One of my final-exam problems in fall 2015 began as follows:

  > In one of the largest and most famous public health experiments ever conducted, in 1954 a randomized controlled trial was run to see whether a vaccine developed by a doctor named Jonas Salk was effective in preventing paralytic

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

> polio. A total of 401,974 children, chosen to be representative of those who might be susceptible to the disease, were randomized to two groups: 200,745 children were injected with a harmless saline solution and the other 201,229 children were injected with Salk's vaccine. ... The results of the trial were as follows: 33 of the 201,229 children who got the vaccine later developed paralytic polio, whereas 115 of the other 200,745 children suffered this fate.

I had obtained the background information for this problem by the usual (and lazy) route of reading about the Salk trial in statistics textbooks. When I finally decided to do a bit of proper scholarship and dig into the literature on the Salk trial, I was amazed to find that the actual experiment was vastly messier than the textbook treatment: **?** tells the true story, in which 623,972 children were actually injected either with vaccine or placebo, "and more than a million others participated as 'observed' controls." Meldrum goes on as follows:

> *The statistical design used in this great experiment was singular, prompting criticism at the time and since. Eighty–four test areas in 11 [U.S.] states used the textbook model:*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

> *in a randomised, blinded design all participating children in the first three grades of school (ages 6–9) received injections of either vaccine or placebo and were observed for evidence of the disease. But 127 test areas in 33 states used an "observed control" design: participating children in the second grade (ages 7–8) received injections of vaccine; no placebo was given, and children in all three grades were then observed for the duration of the polio "season."*

The sample sizes 200,745 and 201,229 appear nowhere in Meldrum's article! To paraphrase XM, in the journey from the actual trial to textbook summaries of it, a set of assumptions was introduced deliberately or accidentally, resulting in a substantial over-simplification of reality.

- The word *valid* with respect to statistical analyses is used frequently in this paper. For example, in Section 1.2 XM say "Meng (1994) obtained some initial theory under this inferential uncongeniality, including conditions for Rubin's MI inference to be *confidence valid*, i.e., the interval estimator has at least the claimed nominal coverage" (italics mine), and in Section 2 the authors offer a simple general

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

recipe: "In general, uncongenality should be regarded as the rule rather than the exception, and a simple confidence-valid procedure to combat any degree of uncongenality is to double Rubin's MI variance estimate." While it's arguably true that failing to cover at the advertised level is worse in confidence interval construction than creating intervals that are (much) wider than necessary to achieve the nominal coverage, I'm uncomfortable with relying only on confidence validity when what John Tukey used to refer to as *robustness of efficiency* — are the intervals indeed wider than they need to be while still hitting the coverage target? — is unaddressed: the phrase "at least the claimed nominal coverage" is equally satisfied at nominal 95% by intervals whose actual coverage is 95.01% and 99.999%, and the latter intervals will of course be substantially wider than the former.

This issue arises again in Section 5.2, where XM say "... in the context of constructing confidence intervals, confidence validity permits the actual coverage to exceed the nominal level (Neyman, 1937), and hence a [variance estimate that's biased high by an unknown amount] is accordingly acceptable." I have great respect for Mr. Neyman and his work — as it happens, he was my statistical grandfather, and (as a graduate student at Berkeley) I had the pleasure of many statistical

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

discussions with him; I'm confident (pun intended) that Mr. Neyman would agree with me that inflated variance estimates are only useful for *a fortiori* arguments of the form "my '95%' confidence interval, based on a positively-biased variance estimate, with coverage at least nominal, doesn't include 0, so the effect I've identified is unlikely to be a statistical artifact." But what can we say if such an interval *does* include 0? XM of course understand this; they conclude Section 5.2 with the statement "... much more research is needed to investigate the general properties of these bounds ..."; hear, hear.

Having grumbled about inflated variance estimates, I'll now hypocritically congratulate XM for having made calculations leading to the simple rule "double T_∞ to yield a nominal 95% interval with actual coverage between 95% and 99.5%," and — if this were a *Royal Statistical Society* Read Paper — it would be my pleasure to either propose or second a vote of thanks.

References

Meldrum M (1998). "A calculated risk": the Salk polio vaccine field trials of 1954. *British Medical Journal*, **317**: 1233–1236.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0474

University of California, Santa Cruz

E-mail: draper@ucsc.edu