Rejoinder

Xiao-Li Meng

I thank *Statistical Science* for organizing a balanced discussion. I feel fortunate that nearly every part of my paper is touched upon by discussants' comments. I will subtitle my responses with one main topic attributed to each discussant; I hope the discussants will agree with my classification. The criticism of the extended combining rules seems universal, and I list the topic under my own name because I want to criticize my rules, too! This Rejoinder is intended as a dessert (or appetizer)—tasteful but light (a bit heavier for Dr. Fay, according to his order).

FAY: FRACTIONALLY WEIGHTED (MULTIPLE) IMPUTATION

"Weighted" implies multiple, so I feel safe that Dr. Fay will not accuse me of putting words (i.e., multiple) into his term; I think it is for the users' benefit to name a method as descriptively as possible. Of course, adding "multiple" after "weighted" is a bit redundant, but I guess Dr. Fay will not mind as modifying "weighted" with "fractionally" also seems redundant.

After a few years of development (documented since 1991), Dr. Fay's proposal now also adopts the framework of multiple imputation. I am particularly pleased to see such a consensus, because now we can concentrate on how to create multiple imputations and how to analyze multiply imputed data. I also hope that this will send a message to those who have been justifying single imputation by citing Dr. Fay's "criticism" of *multiple* imputation. There are nonstatistical arguments for single imputation, such as cost and public perception of authenticity, but it is ironic for any statistician to argue for single imputation. Doctor Fay understands these nonstatistical constraints far better than I do, and yet still adopts multiple imputations. I hope others will not overlook this fact when they cite Dr. Fay's work.

Given the consensus that imputation should be created multiply, Dr. Fay's proposal, at least for now, differs from Rubin's approach in creating imputations and thus accordingly in analyzing the imputed data. Rubin's strategy for creating multiple imputation is to follow the Bayesian recipe—first simulating parameters from their posterior distribution under a specified model, and then simulating missing data from their sampling distribution conditional on the observed data and the simulated parameters; the actual cooking may vary (e.g., the model can be implicit, such as with the Bayesian bootstrap; Rubin, 1987, Chapter 3), but the two steps are necessary in general if the imputations are to reflect fully their uncertainties under a posited imputation model. The advantage of having imputations with the proper variability is that consequent analyses can follow simple laws of probability, such as "total variance equals within-imputation variance plus between-imputation variance." Rubin's approach of analyzing multiply imputed data is then built upon such simplicity—analyzing each imputed data set as if it were real and then combining these "complete-data" analyses by applying the simple combining rules reviewed in Section 2.4 of my paper. The theoretical part of my paper establishes the validity of these rules under conditions that are broader than previously investigated.

In contrast, Dr. Fay's current proposal (1994a) is to skip the first step of Rubin's strategy, that is, the step that reflects uncertainty in estimating model parameters. I have not seen Dr. Fay's general approach for creating his multiple imputations, but from his discussion of the approximate Bayesian bootstrap in Fay (1994a), I surmise that Dr. Fay's strategy is to fix the model parameters at some estimates. His approach of analysis is then to combine these imputations by weighting before performing one analysis—in contrast to Rubin's approach of performing separate analyses and then combining. Since Dr. Fay's imputations do not reflect all of their uncertainties, the simple variance decomposition adopted by Rubin no longer applies, and thus more complicated techniques need to be developed. In other words, Dr. Fay's analysis approach cannot take advantage of the simplicity of computing complete-data variances, but relies on specifically constructed procedures according to his imputation scheme and possibly also the estimators being used.

I do not view the computational complexity of Dr. Fay's analysis procedure as a disadvantage, if that is what it takes to solve the real problem; as I emphasized in Section 1, the focus here is not on computation. What I cannot see is how his approach can better handle the uncongenial cases, where Dr. Fay claims that Rubin's approach does not apply. Take the simple example in Section 3.1. My understanding of Dr. Fay's approach is to fix the subclass (corresponding to $x = 1$) mean at some estimator when cre-

ating imputations, instead of drawing it from some posterior, as in (3.1.3) or (3.1.4). If he fixes it at the subclass sample mean $\bar{y}_{1,!}$, then the issue of uncongeniality disappears. If he fixes it at the overall sample mean $\bar{y}_!$, then his resulting estimator of the subclass mean is essentially identical to my $\bar{\theta}_\infty$, given a reasonable number of imputations, because both estimates are linear functions of the imputed data. Consequently, Dr. Fay must have a powerful method that can compute the variance of $\bar{\theta}_\infty$ without explicitly recognizing the extra information (e.g., that the two subclasses have the same population mean) built into the imputations, and the method must be applicable in general since Dr. Fay's objective apparently is to improve Rubin's approach in general practice.

I am very interested in learning about such a powerful method, but all I can find in Dr. Fay's *Proceedings* papers and technical reports, to this date, are general claims and vague hints with some numerical illustrations that I do not know how to replicate; others have mentioned similar difficulties (e.g., Little, 1991). Doctor Fay suggests that heuristic arguments in statistical papers should be checked via Monte Carlo studies. I completely agree. In fact, I would further suggest that when we make general claims we should first give our heuristic arguments. Furthermore, if possible, we should tell others how exactly a new method should be implemented, so they can replicate it with reasonable effort. Even if one enjoys a jigsaw puzzle, one still has to have all the pieces!

Reading Dr. Fay's comments is just like assembling a jigsaw puzzle with missing pieces. For example, I am puzzled why Dr. Fay cannot see that the purpose of my paper is not to establish the validity of Rubin's approach under congeniality, but rather under uncongeniality. Where in my paper do I give any implication that consistent variance estimates are inappropriate (if these variance estimates are relevant; I'll return to this point later)? Where is an example that Rubin's approach, when implemented correctly, produces a confidence interval that is too short? As I proved in Section 4.4, this can happen only if the imputation model is *information irregular* (given, of course, it is proper), and, as I stated there, I am very interested in finding a real scenario in which this happens. The "binary cluster example" provided in Fay (1991), cited in Dr. Skinner's comments, is highly misleading to say the least. The imputation procedure there was improper, because it (at least) violates condition (ii) of my Definition 3, according to the calculations Fay (1991) provided. Imputation models should never ignore design variables (e.g., clustering), and it is essentially mindless for an imputer to predict $\{0,1\}$ or $\{1,0\}$ clusters when only $\{0,0\}$ and $\{1,1\}$ clusters can be observed and the nonresponse mechanism is assumed to be ignorable. Cooking "Kongbao chicken" without following the clearly written recipe, and then blaming the Chinese recipe for the bad taste does not seem to be a wise strategy for a food critic. Incorrect implementation of Rubin's approach (and other types of problems) was also found in some other examples in Fay (1991); see the discussion by Little (1991).

Returning to Dr. Fay's question about "consistent variances," I would like to add some complimentary oranges to his order. To decorate a Chinese fruit plate, I need 2 pounds of fresh oranges. From a TV commercial, I knew that the best price in the city is 69 cents per pound, but I only had time for a quick shop nearby. The same type of orange was indeed on sale in a local store, slightly more expensive and only in bags—$1.49 a bag with a guaranteed 2-pound minimum. There was another type of less fresh orange that was not on sale—79 cents per pound. The pricing seemed illogical, but anyone who shops knows that paying more does not necessarily mean getting better quality.

The choice was obvious to me, so I grabbed a bag and paid $1.49. To check the store's claim, I weighed the oranges when I got home: 2 pounds and 2 ounces, which means that I almost got the best price in town (*the ideal procedure*)! I mentioned my "bargain" when I presented Dr. Fay with the fruit plate. To my surprise, Dr. Fay responded that I should have bought the more expensive type, because then I would have gotten exactly 2 pounds (*confidence coverage*) as I intended. "But," I was obviously puzzled, "I got more than I wanted with less money (*length of a confidence interval*), and I got the fresh ones (*efficiency of the estimators*)! Besides, even if I had bought the other type, it is very unlikely that the weight would have been exactly 2 pounds since the store does not sell oranges by the slice (*exact confidence coverage is rare in reality*), so if I really wanted 2 pounds, it is very likely that I would have had to buy more than 2 pounds. In fact, I was glad that I had extra because I was so excited about the bargain that I accidentally dropped a couple of slices when decorating the plate (*errors in approximations*). But with the extra, I was able to cover the plate (*slightly conservative confidence intervals are preferred when approximations are made*)!" I was quite confident that I must have convinced Dr. Fay until I heard his question: " Does your complex argument lead us to a conclusion that, if you bought more than 2 pounds, buying exactly 2 pounds is inappropriate?"

Some readers may feel my "fruitful" metaphor unnecessary, but I want to make sure these points are clear, because Dr. Fay has already complained that my argument was too complex and requires "a

large number of concepts, terms and so on." To be fair, Dr. Fay's point is not the most unusual one that I have seen; at least both Dr. Fay and I agree that we want oranges (*sensible inferences for meaningful estimands*). I have seen arguments, even in very recent articles, to the effect that: "I want exactly 2 pounds. The actual fruit is only a matter of taste."

I am also quite puzzled by Dr. Fay's implication that my paper is irrelevant for complex survey designs. As I explicitly stated in the paragraph after the Main Result (Section 4.4):

> In addition, since no condition imposes any restrictions on survey designs or nonresponse mechanisms, all of the results are completely general and even apply to cases in which our simplified notation is inappropriate (e.g., sequential surveys, unstable response; ...).

The reason that I can make such a strong statement is that my investigation adopts the framework of efficiency, which is universal to any design or model, just as the notion of variance is universal to any estimator. My results only involve basic concepts of statistics, proved with elementary mathematical statistics, peer-reviewed by at least four reviewers (one of them even suggested that I drop the Appendix because it is too short) and can be checked by any reader without Monte Carlo studies or desktop computers. If there was anything wrong with my results, I assume Dr. Fay would have pointed it out. Large-scale surveys are always complex, and I completely agree with and appreciate Dr. Fay's emphasis on complex samples, especially considering the problems he needs to deal with at the Census Bureau. However, I am puzzled that my statement quoted above, which was made specifically to emphasize the applicability of my results to complex surveys, did not catch Dr. Fay's attention as well as some other parts of my paper did. Speaking of Dr. Fay's role in the Census Bureau, I also find it a bit odd that he implies, seemingly in response to my comments at the end of Section 1.2, that I put too much emphasis on public-use files. My emphasis there is that Rubin's approach, compared to Fay's proposal, is much more advantageous when data files are shared by multiple users for various analyses. (Is there any survey listed in Dr. Fay's comments that does not fall into this category?) I am disappointed that I cannot find a clear answer in Dr. Fay's discussions to my question raised there.

Let me finish my serving to Dr. Fay by returning to his *fractionally weighted (multiple) imputation*. In the early stage of developing "a general and flexible system for handling nonresponse in sample surveys," Rubin (1977; the quote is essentially its title) considered three ways of representing a distribution of imputed values, as summarized in three of his subsection titles:

> 2.1. Formally and analytically finding predictive distribution for summary statistics;
>
> 2.2. Simulate full predictive distribution via many imputed data sets;
>
> 2.3. Perform one weighted analysis with several versions of missing data.

Of these, Section 2.2 is essentially the birth of Rubin's approach, with Section 2.1 being its theoretic parents. Section 2.3 was a little brother that eventually disappeared (abandoned?) from the family, but whose descriptions seem quite close to the descriptions provided for *fractionally weighted (multiple) imputation*. Could it be that the little boy has been found after so many years (obviously grown up)? Since Rubin (1977) is not easily accessible to general readers, I list the first paragraph of Rubin's Section 2.3:

> For some analyses, it may be possible to obtain information similar to that described in Section 2.2 without having to repeat the same analysis several times. The idea is to use the same type of filled-in data set as described above, but now we would include all five[1] versions of each unit in one analysis: the weight for each unit with missing data would be split among the five versions of his[2] filled-in data, and a weighted analysis would be done.

In case the little boy disappears again, I also quote from Fay (1994a), which was presented at the 1993 ASA meeting:

> The FWI[3] estimator assigns a fractional weight to each imputed value. For example, if the analysis of a complete data set would be unweighted, then each of the m imputed values should receive the weight $1/m$, and observed values receive a weight of 1. More generally, the m imputed values should divide the original weight for the case equally. This approach represents a fundamental difference from MI[4]: for FWI, a single, weighted analysis of the data set is

[1] In his Section 2.2, Rubin illustrated with five imputations (i.e., $m = 5$).

[2] Refers to the analyst mentioned in Rubin's Section 2.2.

[3] Fractionally weighted imputation.

[4] Multiple imputation.

envisioned, instead of m separate analyses in (2.1).[5]

I make such a comparison to speculate that Dr. Fay's approach may eventually evolve into one that is very similar to Rubin's, in view of the evolution of Rubin's approach. In fact, the idea of "weighted imputations" is an old one, attributed by Rubin (1977) to Beale and Little (1975) and Little (1976; eventually published as Little, 1979). Doctor Fay's list of his current and planned work provides further evidence for my speculation—using models and handling "respondent missingness" (nonresponse mechanisms?) are already routine components of Rubin's approach.

Finally, for the record, if Dr. Fay's approach eventually evolves into one that is better than Rubin's, which is also evolving, I will be among the first to abandon Rubin's approach partially or entirely according to the extent of the superiority of Dr. Fay's approach as evidenced in routine applications. It is my belief that the ultimate task for a statistical researcher is to advance our ability to solve applied statistical problems (the adjective "applied" is redundant but I follow the convention), but the unwillingness to abandon one's own investment, admittedly a painful process, often prevents us from accomplishing such a task.

SKINNER: SELF-EFFICIENT ("ANTIBOOTSTRAP") ESTIMATION PROCEDURES

I thank Dr. Skinner for finding that my framework is helpful and for making specific comments. Doctor Skinner is mainly concerned with "whether the conditions of the Main Result are always reasonable, in particular the assumption 'most of the time analysts will use (asymptotically) efficient estimators.' " Two modifications need to be made here before I discuss Dr. Skinner's specific example. First, no mathematical condition is always reasonable, otherwise it would not be called a condition. Second, "most of the time..." is not an assumption of my results. My results assume that the analysts use *self-efficient* estimators, an assumption that is much weaker than *efficient* estimators. The latter implies the former, and the quoted statement appears after the Main Result to emphasize that the condition of self-efficiency is really not a restriction in practice given the type of estimators analysts are currently using. I am glad that Dr. Skinner mentioned cluster sampling, since this is a design that Dr. Fay has repeatedly claimed that Rubin's approach cannot handle, and I will show that self-efficiency holds easily in general practice for cluster estimators.

To be consistent with the notation in Section 4.3, I relabel Dr. Skinner's estimators with $\hat{\theta}_0 = \bar{y}$ and $\hat{\theta}_1 = \sum R_i y_i / \sum R_i m_i$, and write θ as the population mean being estimated. The response mechanism Dr. Skinner illustrated can be viewed as a second-stage cluster sampling with a random sample size $n_1 = \sum R_i$ from the n originally sampled clusters. To simplify illustration, I will condition on n_1 in the following calculations; otherwise one has to deal with the possibility that $n_1 = 0$, which implies that the mean-squared error of $\hat{\theta}_1$ is not well defined, a theoretical complication that is irrelevant here.

Now, for any constant λ, it is easy to check that

$$
\begin{aligned}
&E\left[\left(\lambda\hat{\theta}_1 + (1-\lambda)\hat{\theta}_0\right) - \theta\right]^2 \\
&\quad = \lambda^2 E[\hat{\theta}_1 - \hat{\theta}_0]^2 + E[\hat{\theta}_0 - \theta]^2 \\
&\qquad + 2\lambda E\left[(\hat{\theta}_1 - \hat{\theta}_0)(\hat{\theta}_0 - \theta)\right],
\end{aligned} \tag{1}
$$

where the expectation is taken over the two-stage cluster sampling with the sample sizes fixed. If all the clusters are of equal size, as in Fay's example cited by Dr. Skinner, then $\hat{\theta}_1$ is unbiased for $\hat{\theta}_0$ when conditioning on the first-stage sampling. Thus, the cross term in (1) is zero, implying that any mixture of $\hat{\theta}_1$ with $\hat{\theta}_0$ can only have higher mean-squared error than $\hat{\theta}_0$ itself. Thus, contrary to Dr. Skinner's intuition, $\hat{\theta}_0 = \bar{y}$ is self-efficient.

When the clusters are not of equal sizes, the biases caused by ratio estimators can make the cross term nonzero. But then the cross term is of the order $(r^{-1} - 1)n^{-2}$, while the other two terms on the right-hand side of (1) are of order $(r^{-1} - 1)n^{-1}$ and n^{-1}, respectively, where $r = n_1/n$ is the response rate. We all know that in typical survey analyses terms beyond the first order are considered practically irrelevant because n is large. When n is small (as with small-area estimations) or the response rate is low, the above calculations can be deceptive. But then practitioners have been warned to stay away from ratio estimators if their biases are no longer negligible, and to replace them with more sophisticated ones such as jackknife or Bayesian estimators. Doctor Skinner surely understands that whenever a delta method is employed for computing variances it is assumed that the biases in the corresponding estimators have been reduced to a negligible degree. Once the bias is negligible, so is any violation of self-efficiency, as illustrated above.

In summary, *self-efficiency* is a very weak requirement. It can be defined purely in terms of randomization distributions for a given estimation procedure, so it has little to do with model misspecification, con-

[5]Equivalent to (2.4.2) of my paper.

geniality or even the (asymptotic) efficiency we are familiar with. It has much to do with "unbiasedness," a concept that is dear to the heart of traditional survey practitioners. Doctor Skinner seems to have been misled by Fay's (1991) deceptive example, which I discussed earlier, and/or by my choice of the term (I apologize for the latter). Doctor Skinner's intuition appears to come from comparing different estimation procedures, as we often do when we talk about efficiency, but my definition of self-efficiency is only concerned with a given procedure. Perhaps I should have adopted the term "antibootstrap" (suggested by Professor Raghu Bahadur) or "antijackknife," but I wonder if that could cause other confusions; I welcome any suggestion.

Regarding Dr. Skinner's comment on the "powerful feature," he is technically correct if single imputation is defined as one draw from a missing-data predictive distribution (i.e., Rubin's approach with $m = 1$); a single data point is indeed unbiased for its mean, but I do not see why this needs to be emphasized. Given various ways that have been suggested for single imputation, and now even for multiple imputation, I suggest that, whenever possible, we specify its content when we mention single or multiple imputation. The readers may have noticed that I use "Rubin's approach" instead of "multiple imputation" in the rejoinder.

I would like to finish my serving to Dr. Skinner by thanking him for his valuable support of using Bayesian prediction for imputation—I have no disagreement that other approaches, used separately or together with the Bayesian approach, may be more cost effective under special circumstances; Belin et al. (1993) is in fact a good example, as discussed in Zaslavsky's comments. I would also like to present, for entertainment purposes only, Dr. Skinner with a fortune cookie containing two trivia questions: Who claimed that biases can be removed entirely in reality, and who suggested that imputed values should not be flagged?

SCHAFER: THE COMPLEXITY OF IMPUTATION MODELS

Schafer's comments are particularly valuable, since he has substantial first-hand experiences in creating multiple imputations with large-scale data sets. "Schafer's algorithm," as detailed in his forthcoming book (Schafer, 1994), has already been frequently requested by multiple imputers, and I believe it will soon be recognized as one standard algorithm in statistics for handling multivariate incomplete data. I thus feel particularly encouraged by Schafer's appreciation of my technical work and my choice of the yardstick.

Schafer's comments on the complexity of imputation models address a well-known issue in practice: desirability and feasibility may not go together. Schafer's *model trimming* provides a very useful and specific strategy for achieving the goal I discussed in section 6.1: "...the imputation model should be as objective and general as the imputer's resources allow." With model trimming, one starts with a desired general and saturated model, and then gradually introduces model assumptions to allow straightforward estimation. I prefer this strategy because it allows the imputer to see clearly what has been sacrificed during the trimming, and thus to anticipate the type of uncongenial analyses and to warn the users if necessary.

I want to emphasize, however, that Schafer's concerns arise because of the complexity of the underlying problem, not of Rubin's approach. Such distinctions are sometimes not well understood, such as blaming the EM algorithm for the multimodality of a likelihood (see the Rejoinder to Meng and Rubin, 1992a). I want to make this point crystal clear so Schafer's concerns will not be (mis)cited by others as arguments against Bayesian imputation. There is simply no free lunch. The free lunch for a user (in terms of handling nonresponse) comes at the imputer's expense, which is substantial when many lunches need to be paid for. Reductions in the expenses can only decrease the quality of the lunch, unless we can find a whole new way of preparing it.

Regarding Schafer's comments on the number of imputations, I surely agree that whenever possible one should use the maximum number of imputations available. Here, the problem is simply one of choosing simulation sizes—the more the better, as long as one can afford them. In my own studies, such as the one that I will describe briefly in my response to Zaslavsky, I have also found that significance levels generally become quite stable when $m \geq 10$, and I have used $m = 50$ (much larger than necessary), essentially removing the noise caused by finite m. I suggested allowing users to select randomly, simply because I do not want to create the impression that just because 30 imputations are provided they all have to be used before valid conclusions can be reached.

Finally, I greatly appreciate Schafer's support for creating a generous number of imputations (e.g., $m = 30$) with public-use data files. Of course, great efforts are needed before this can become a common practice. Schafer, together with others, has been quite successful in bringing reasonably large m (e.g., 10) into some large-scale data files (e.g., Schafer, Khare and Ezzati-Rice, 1993). For that, I finish my serving to Schafer with a full glass of champagne: Congratulations and more power to you!

ZASLAVSKY: HYBRID INFERENCES

I apologize to Zaslavsky for making him wait for so long, but he has been my guest so often that I have difficulty in finding a dessert that he has not tasted. I only wish that I could have his delicious recipes on my menu, such as his wonderful summary of hybrid inferences in my Section 2.2.

I do have an example to add to Zaslavsky's examples on using the hybrid approach to handle complex problems. This one is on estimating survival after an AIDS diagnosis, from the surveillance data maintained by the Centers for Disease Control (CDC). As in all of Zaslavsky's examples, the problem we (Tu, Meng and Pagano 1993a, b) faced was complex; our *JASA* article lists the difficult issues, mainly caused by the complex nature of the CDC data. One key difficulty was that reporting of deaths was delayed and the time of reporting was not available for each death. A joint likelihood of time of AIDS diagnosis, time of death and time of reporting of death is in principle possible, but very difficult to implement. We thus adopted a hybrid approach: we first used whatever was available to construct a Bayesian model to multiply impute the number of delayed cases, and then we used a discrete proportional hazard model to fit a survival distribution to each imputed data set. Finally, we followed the standard combining rules to reach our conclusions. Our imputation model and our analysis model were obviously uncongenial (although at that time I had not defined such a notion), not only because they were built upon different amounts of data, but also because modeling reporting behavior is a different objective from modeling an AIDS survival distribution.

Were we sure that the results from our hybrid approach were correct? I am pretty sure that they were incorrect if the "correct" answer existed. Since we were dealing with data collected from real life, there is no "objective" Monte Carlo study that can confirm our results. However, a bright side of dealing with real data is that we can check the plausibility of our results against reality. Such checking does not necessarily validate a statistical conclusion but can invalidate one. For example, our results confirmed an unfortunate reality—statistically, nearly no patient with *Pneumocystis carinii* pneumonia (PCP, an AIDS-defining diagnosis) could survive longer than five years after the diagnosis. Some other analyses, discussed in our *JASA* article, predicted much higher five-year survival rates (e.g., 15%), a clear statistical miracle when compared to reality.

Given the flexibility and power of the hybrid approach for handling complex problems, we surely all want to see more theoretical studies as well as empirical evaluations. Combining different perspectives is also a great stimulus for theoretical work, at least in my experience. I would like to list several other "hybrid" works that I am involved in, to advertise the opportunities of theoretical research in this exciting "hybrid" field: posterior predictive p-values (Meng, 1994); Bayesian model checking using tail area probability (Gelman, Meng and Stern, 1995); and single observation unbiased prior (SOUP) (Meng and Zaslavsky, 1994).

I would like to finish my serving for Zaslavsky with the SOUP that we have been cooking together for quite a while, especially because this work arose from a multiple-imputation problem at the Census Bureau, and it relates to Zaslavsky's point on the frequentist's emphasis on the unbiasedness of point estimators and the Bayesian's task of specifying probability models. When we construct a Bayesian model for imputing the number of uncounted households in a census block, we would like to make sure that the imputed total is unbiased for the actual total in that block when averaging over the sampling and imputation distributions. The unbiasedness here is desirable because the imputed totals are often utilized under aggregations over many census blocks. We thus faced a practical as well as theoretical problem—what kinds of priors will yield a posterior mean that is identical to a desirable unbiased estimator? Our work answers such questions for many common statistical models.

MENG: THE ATTEMPT TO EXTEND RUBIN'S RULES

Both Skinner and Schafer expressed concern about the extended combining rules based on importance sampling. Their criticisms are right on target. For the record, I want to announce even more loudly (see Section 5.1) that the extended rules were only proposed for "brain-storming" purposes, as my entire work presented here essentially started with that thought. It may be a long journey before we can find a workable procedure for dealing with the cases that my extended rules targeted, but I am not too concerned with this because so far these cases only arise in artificial examples designed to challenge Rubin's rules.

I do think, however, that there are several important messages we can learn from the extended rules. I illustrate them by briefly mentioning how my thoughts evolved during the work reported here. I had been interested in the "uncongenial" problem for quite a while, but I had not seriously considered the problem until I read Kott's (1992) "counter-example," a recast of one of Fay's (1991) examples, and essentially the same example used in Section 3.1.

My first reaction was that the "counter-example" must be caused by "uncongeniality," because otherwise it is impossible to have a discrepancy since Rubin's rules follow the laws of probability, as discussed in Little (1991). Once the uncongeniality was found, I thought about possible ways to adjust it, and the idea of importance sampling came naturally. But then the most important message arose: the extended rules make adjustments by first adjusting the estimator itself. Much emphasis in Dr. Fay's work has been on the consistency of variance estimators, as my "orange" example illustrates. However, if the uncongeniality leads to inconsistent estimators, then only fixing variance estimators solves the wrong problem. In this regard, I will maintain my extended rules as a yardstick for future more workable procedures.

In order to compute the weights required by my extended rules, I need the imputation densities defined in (5.3.1). This reminded me: besides the imputations, there are other useful imputational quantities that the imputer (e.g., Census Bureau) can provide without compromising any confidentiality. In my first submission (Meng, 1993), which mainly focused on the extended rules, I also discussed the possibility of having the imputer provide the imputation densities for some common summary statistics, which partially addresses Dr. Skinner's concerns. It is quite clear that the more the imputer can provide, the better the analyst can do; I am very much looking forward to research in this direction.

A third message arose when I realized that my extended rules, in their current form, essentially wipe out all of the imputer's effort, as I discussed in Meng (1993). This is very stupid! (The discussants may be too polite to say so even if they wanted to, but I guess I am allowed to say that to myself.) Even if there is some undesirable uncongeniality due to deficiency in a particular aspect (e.g., a particular dependence was incorrectly specified) of an imputation model, the chances are that the imputer's model, as a whole, is still far more sensible, for the purpose of imputation, than the analyst's model with the correct specification on that particular aspect. The message here, then, is to make only a partial adjustment to maintain the correct (majority) part of the original imputations. Fay (1994b) seems to be hinting at such a direction, and I am looking forward to the details. I have been discussing such partial adjustments with one of my students, but I am unsure if we can come up with anything that will be as simple and general as Rubin's approach and yet will outperform it. I have already failed once!

In short, my whole attempt to extend Rubin's rules ended up with confirming and establishing the remarkable robustness of his rules. I know most of us want simple rules that work. For that, I pour myself a cup of "House Wine of Confucius" (my favorite) to celebrate my failure!

IMPUTATION SUPERHIGHWAY

My final comment is reserved for Dr. Fay's Figure 1, whose "high-way" and "low-way" inspired the following analogy, perhaps with a little imagination. Both Fay and Rubin are asked to handle the problem of road damage caused by an earthquake. Rubin's "high-way" approach is to repair the damaged roads as much as possible, and thus a driver can follow his familiar route to his destination. Dr. Fay, however, seems to suggest, as I surmised from his "low-way" paradigm in the figure, to ask the driver to use the undamaged roads to reach a different destination, and then somehow (build a new highway?) to allow him to drive (or give him a ride?) from the new destination to his desired destination.

As I said, I am looking forward to Dr. Fay's "low-way" approach. But until I am given explicit instructions on how to drive on his low-way, I will only drive on Rubin's imputation superhighway, because I want to know where I am heading when I am driving!

(Sorry! No drink here. We are on a highway!)

ADDITIONAL REFERENCES

BEALE, E. M. L. and LITTLE, R. J. A. (1975). Missing values in multivariate analysis. *J. Roy. Statist. Soc. Ser. B* **37** 129–145.

CITRO, C. F. and HANUSHEK, E. A., eds. (1991). *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling* **1, 2**. National Academy Press, Washington, D.C.

CITRO, C. F. and HANUSHEK, E. A. (1994). Estimating the effects of proposed legislation: the case for model validation. *Chance* **7** 31–40.

FAY, R. E. (1994a). Valid inferences from imputed survey data. Unpublished manuscript.

FAY, R. E. (1994b). Inferences from survey data employing mass imputations. Prepared for presentation at the Annual Meetings of the American Statistical Association, Toronto.

GELMAN, A., MENG, X. L. and STERN, H. (1995). Bayesian model checking using tail area probabilities. *Statist. Sinica*. To appear.

KALTON, G. and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology* **12** 1–16.

KUHN, T. (1962). *The Structure of Scientific Revolutions*. Univ. Chicago Press.

LITTLE, R. J. A. (1979). Maximum likelihood inferences for multiple regressions with missing values: a simulation study. *J. Roy. Statist. Soc. Ser. B* **41** 76–87.

LITTLE, R. J. A. (1991). Discussion of "A design-based perspective on missing data variance" by R. E. Fay. In *Proceedings of the 1991 Annual Research Conference* 441–446. U.S. Bureau of the Census, Washington, D.C.

LITTLE, R. J. A. (1993). Discussion of "Hierarchical logistics regression models for imputation of unresolved enumeration status in undercount estimation," by T. R. Belin, G. J. Diffendal, S. Mack, D. B. Rubin, J. L. Schafer and A. M. Zaslavsky. *J. Amer. Statist. Assoc.* **88** 1160–1161.

MENG, X. L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 000–000.

MENG, X. L. and RUBIN, D. B. (1992a). Recent extensions to the EM algorithm (with discussion). In *Bayesian Statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) **4** 307–320. Oxford Univ. Press.

MENG, X. L. and ZASLAVSKY, A. M. (1994). Single observation unbiased prior. Technical Report 393, Dept. Statistics, Univ. Chicago.

MULRY, M. H. and SPENCER, B. D. (1991). Total error in PES estimates of population. *J. Amer. Statist. Assoc.* **86** 839–855.

MULRY, M. H. and SPENCER, B. D. (1993). Accuracy of the 1990 census and undercount adjustments. *J. Amer. Statist. Assoc.* **88** 1080–1091.

RUBIN, D. B. (1977). The design of a general and flexible system for handling non-response in sample survey. Project report for Social Security Administration.

SÄRNDAL, C. E., SWENSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling.* Springer, New York.

SCHAFER, J. L. (1994). *Analysis of Incomplete Multivariate Data by Simulation.* Chapman and Hall, New York. To appear.

SCHAFER, J. L., KHARE, M. and EZZATI-RICE, T. (1993). Multiple imputation of missing data in NHANES III. In *Proceedings of the Annual Research Conference* 459–487. U.S. Bureau of the Census, Washington, D.C.

SCOTT, A. and SMITH, T. M. F. (1969). Estimation in multistage surveys, *J. Amer. Statist. Assoc.* **64** 830–840.

SKINNER, C. J., HOLT, D. and SMITH, T. M. F. (1989). *Analysis of Complex Surveys.* Wiley, New York.

ZASLAVSKY, A. M. (1991). Estimation of bias and variance in the dual system estimator. Unpublished memorandum.

ZASLAVSKY, A. M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *J. Amer. Statist. Assoc.* **88** 1092–1105.

ZASLAVSKY, A. M. and THURSTON, S. W. (1994). Error analysis of food stamp microsimulation models. In *Proceedings of the Section on Survey Research Methods.* Amer. Statist. Assoc., Alexandria, VA. To appear.