DISCUSSION

Walt R. Simmons, National Center for Health Statistics

1. Comment on Fellegi and Gray Paper.

1.1 Let me say first that this paper was a pleasure to read for at least three reasons:

A. It presents a *simple* and *operational* procedure for calculating the approximate sampling variance of a statistic from a complex survey, without going into a welter of the complicating factors which are always present, and which worry many of us who deal with these matters.

B. I'm especially pleased to see the procedure designed and used not alone for securing the standard error of a statistic to be published, but as an organized system of operational control of the survey. This reviewer believes that one of the most serious weaknesses in survey work today is the failure of survey execution to mirror faithfully the design. The control system described by Fellegi and Gray, based on the variance analysis patterns is good medicine for this ailment.

C. I give a cheer for the recognition that there is substantive information in the time series of variances compared with the simple random model variances that goes beyond the usual measure of precision of the sample statistics. I agree fully that this is a real contribution to the information provided to the analyst and economist.

1.2 Let me turn, then, to a few criticisms and questions.

A. I found the notation and presentation in the Appendix less clear or clean than it might have been. This began with a couple of typos in my draft, and with a very skimpy definition of the "balancing unit factor" B_{jb} at the beginning. It wasn't helped by the choice of B for both "balancing factor" and "variance," nor the oddly worded introduction of the symbol h.

B. A more substantial issue is closely allied to one of the strengths of the recommended procedure. The estimates of variance themselves have a sampling error, of course. Thus observed differences between two variances may represent real differences, or they may reflect only random variation. Analysts need be warned of this circumstance, lest they draw unwarranted conclusions; just as they might from looking at primary statistics, and disregarding their standard errors. Perhaps there should be advice to analysts to look at *patterns* of variances rather than a simple variance or pair of variances in drawing conclusions.

C. This variance of estimated variances raises another point which is perhaps a matter of taste. If one plots estimated relative standard errors against a series of statistics from a multi-purpose survey, he finds generally that the former decreases in a fairly smooth pattern with increases in the latter. There are exceptions or deviations from a smooth curve. Some are real, some are themselves random variation. It seems to me that when the statistical agency publishes a very large number of estimated variances, not all of which will have an internal consistency, the consumer is puzzled. I lean in favor of publishing a more limited number of average or "typical" variances or relative standard errors. This point has importance, too, when one is focusing attention on the ratio of estimated variance to the variance of the simple random model. If this point of view is not accepted, I would argue that seven levels of coded published variances are too many. It is too many for the consumer to keep in mind, and too many because it overstates the precision with which the variances are estimated. Might not this be better:

a for $CV < 5\%$

b for $5\% < CV < 15\%$

c for $CV \geqslant 15\%$

1.3 The Canadian linearized variance procedure is efficient for estimating variances of means, difference of means, and simple ratios. This paper does not discuss the problem of estimating variance of more elaborate statistics such as regression coefficients, or of position statistics. There another technique—perhaps balanced pseudo-replication—is needed. Pseudo-replication is useful, too, as a device for discovering the relative impact on overall variance of different design features and different estimator features.

1.4 May I say again, I liked this paper.

2. Comment on Frankel Paper.

2.1 The Frankel paper is in my judgment one of the most significant and satisfying pieces of research to come out in recent years in the realm of applied survey sampling. On the analytic side, in survey work, our ultimate objective is usually to estimate from sample data a first order statistic, say θ', of the parameter θ and then a standard error S_θ of θ', form the ratio $\frac{\theta'}{S_{\theta'}}$, *assume* that the ratio is distributed as t- or normal, about a mean of $\frac{\theta}{S_{\theta'}}$ with unit variance. When the sampling is simple-random, and θ is, say a mean, both theory and empirical evidence have justified this approach. When the survey design is more complex, involving ratios, clustering, stages, phases, post-stratification—and θ is a more elaborate parameter such as a median, ratio, or correlation coefficient, we have not been sure how to calculate $S_{\theta'}$ and have been quite unsure of the real distribution $\theta/S_{\theta'}$. Frankel's paper, and his dissertation behind the paper, have taken a long and welcome step toward resolution of both problems.

2.2 All survey samplers should be grateful for these results. I'm especially pleased, perhaps, because the results confirm the practice that is being followed today in the National Center for Health Statistics. We use both the linear scheme and BRR—which we call Pseudo-replication, and which we prefer for its analytic capabilities. The BRR approach was evolved by Philip McCarthy, building on the work of others while he was searching for "methods for analyzing data from complex surveys," working under an NCHS contract.

2.3 Frankel did find BRR "best" under his (I think, appropriate) criterion in every one of 90 plus estimates tested. But he found, also, that both the Taylor expansion technique and Jackknife gave acceptable results. For simple statistics such as means, or ratios, or their differences, he endorses the linear approach because it is cheaper. For more elaborate statistics, he chooses BRR. That seems to be a sound course.

3. Comment on Banks and Shapiro Paper.

3.1 To me, the most significant feature of the paper is the way in which it extends a Census Bureau tradition, and exhibits not merely how a sampling variance of a statistic can be estimated, but how that estimate or estimating process can be further analyzed to tell more about the survey design. Recall that their paper explores the Design Effect; the contribution of between-PSU and within-PSU components; the "bias" of between-PSU estimates of variance in the usual collapsed stratum operation; the impact of the rotation patterns of CPS; the impact of the composite estimator; comparative estimators: linear and replication; and variance of the variance estimators. All of these matters—and others—are important, and deserve attention.

3.2 Banks and Shapiro say that the linear scheme and replication give essentially the same result—i.e. they have empirically the same expected value. I would not quarrel with their conclusion; indeed, my own experience tends to support the declaration. But their evidence in the paper for the statement is not overly strong. For example, for what should be the best estimate (total labor force), the two estimated rel-variances differ by 34% of the smaller one (about 1.5 times the estimated standard error of that difference.) This isn't entirely comforting.

3.3 The authors conclude that the linear scheme provides much more reliable estimates than the replication because typically, the former has a standard error of the order of ½ that of the latter. The evidence on this point is the consequence of a good many factors, some of which I'm in no position to assess. But I would note two:

A. The replication variability would be reduced if more replicates were used—for example, if 40 rather than 20 were used, variance of the variance would be greatly reduced.

B. What effect on estimating variance of the variance in the linear case does the dropping of 2nd order terms have?

4. Comment on Folsom, Bayless and Shah Paper.

4.1 First, I apologize for not having had time to study this paper as carefully as the other three. In particular, I have not attempted to verify the rather extensive mathematics in the paper.

4.2 If, indeed, they have developed a technique for producing unbiased estimates of the contribution to sampling variance from each of three stages of sampling in a complex design, that is a very definite contribution to the design of any similar subsequent survey.

4.3 I'm unable to express a judgment on the relative validity or impact on precision of several factors in the development of this paper, but I might call attention to three which may be significant.

A. Although the authors declare that the method is applicable to any number of sampling units at any stage, a part of the development depends on sampling with replacement in the 1st stage, and an assumption that at least 2 PSU's appear in each stratum, with no PSU appearing more than once in the sample.

B. This study, like some others, appears to secure results from Jack-knifing and from Taylor expansion that are clearly similar. Indeed, the degree of similarity (practical identity) is surprising, considering the quite complex algebraic formulations of both methods. This is particulary notable, since the estimated relative standard errors of the estimated variances are quite substantial, running for the components mostly from 50 percent upward. Is there some fictitious element or redundancy in computation which makes results from the two methods more alike in the numerical example given than might usually be the case?

C. The complexity of the approach suggests that the required computer programming likely is also complex—expecially when non-response is taken into account—and so one must be careful to be watchful for risk of error in this direction.

4.4 It's no detraction from this paper to note that effective design use of variance components depends not only upon knowledge of these components but, also, upon good unit cost data. And the latter are often not known with precision.

5. Summary Comments.

5.1 In summarizing my thoughts on these four papers, I can say as I did in beginning: this is a strong set of papers which I enjoyed and which make real contributions. Collectively they solve, or take significant steps toward solution of a considerable variety of important survey problems. I'm certainly not going to try at this point to catalog those problems or solutions, but may I remind the audience of several important features under discussion today.

5.2 Of first importance is the putting on a sounder basis the drawing of inferences from complex surveys--and as the current saying goes, "that's what it is all about."

5.3 The papers shed a good bit of light on what are today the three leading methods of calculating sampling variance for statistics from complex surveys, and on the comparative advantages of each.

5.4 The authors offer a number of examples and a wealth of leads in showing how analysis of variance estimates can produce added information about sampled universes beyond that commonly obtained in the first order statistic.

5.5 Lest we be too complacent about our successes in these areas, may I note that none of the papers today dealt explicitly with that other fundamental problem in survey work: measurement error—though the proposed estimators include a part of measurement variance.

5.6 Finally, I should like to take advantage of having the floor to point out one of the special advantages of the replication schemes. This is a feature which I and associates at the National Center for Health Statistics have described in earlier papers, and which we find useful and convenient. In using half-sample pseudo-replication, we print out, say, a 2-way table from not only the parent sample but also the same table as estimated from each of the perhaps 20 half samples. Thus for *any* statistic, such as a median or difference between domain means, we can compute a variance on a desk calculator, with a simple computation, using the 20 replicate estimates. This course avoids the necessity of foreseeing every need beforehand, or of a possibly substantial reprogramming of the variance run. It is, if you like, a "Jack-knife of a different color."