DISCUSSION

William D. Kalsbeek, University of North Carolina,
Department of Biostatistics, Rosenau Hall, Chapel Hill, NC 27599

Sampling design in one-time surveys and the initial design or redesign in ongoing surveys brings to light many diverse issues. The researcher must contend with these issues in a way that is consistent with the dictates of good science yet respectful of the limitations brought about by timetable and resources. Tradeoffs must be found to balance these often opposing influences. The authors of the papers we have heard here this afternoon are to be congratulated for helping us to see how constraints in the setting within which designs are developed provides fertile ground for developing creative ideas. My remarks are intended to highlight some of the scientific give and take we have seen illustrated in this session.

The solution developed in the Choudhry paper is borne out of the classic confrontation between the mean squared error of survey estimates and the cost of producing these estimates. On the one hand, larger sample sizes leading to smaller variances are affordable when interviews are obtained exclusively by relatively less expensive telephone interviewing. On the other hand, coverage error in a relatively more costly sample interviewed face to face will be lower since those with and without telephones tend to differ. Because both design options have merit, a logical choice is to exploit the benefits of each by using both. This dual frame approach, however, raises the primary issue of the paper; and that is, how many of each type of interview do we take? The answer, as we often see in cost-variance optimization of this type, is that "it depends," although in predictable ways. For example, the less telephone interviews costs and the more efficient their sampling designs are (relative to the face-to-face alternative), the more of them you do and the more heavily you weight their data from the overlap frame in analysis. The illustration given in the paper for the matter of estimating unemployment rates provides a useful gauge as to the actual allocation one should when these estimates are needed.

Three additional features of the computational analyses might have broadened the utility of the findings. First, proportions substantially greater than those of the order of 0.05-0.15 for unemployment rates are often estimated in practice. For larger proportions the ratios of these proportions in the overlap and nonoverlap groups (R_a/R_B) may more commonly be smaller than those used in the paper (1.5 was the smallest value used). Therefore, perhaps it would have been useful to have presumed larger values of R_B and assumed smaller values of R_a/R_B to go with them. Second, with the rising costs of face-to-face interviewing, there is greater pressure to limit the number of PSUs in area samples. This leads to fewer PSUs (to limit travel costs), even for large samples, and thereby larger design effects for some estimates. The range of assumed design effect ratios might have therefore spanned somewhat beyond the upper limit of 1.5 that was used (e.g., to 2.0 or 2.5 perhaps). Finally, the premise of the optimization model is that the ratio of the design effects and the ratio of the sampling rates are functionally unrelated, when in fact the latter would affect the former if the number of PSUs in the cluster sample presumed for the area frame were fixed. An increase in the sample size would increase average sample cluster size which in turn would alter the ratio of design effects. It would have been helpful for the implications of this interrelationship to be incorporated into the study.

The other basic methods paper by Williams deals with another common confrontation between statistical and practical efficacy brought about by the need to introduce complexity into the sampling process through unequal probability sampling of clusters of population elements in order to reduce survey costs. The design thus created presents the survey analyst with unique difficulties in performing analyses that are directed by survey objectives. Whereas substantial effort has been aimed at parametric methods of inference, relatively little has been done to bridge the gap between nonparametric methods and complex survey design. This paper examines some basic distributional properties of estimation via U-statistics from finite populations and considers its application to a rejective technique for PPS selection of PSUs in cluster samples.

Findings from the numerical simulation appear to support the theoretical conclusions regarding the asymptotic distribution of this class of estimators but not to the extent perhaps that the author would have hoped. These findings might have been more confirmatory if sample sizes had been larger than 100, although this is clearly more difficult in simulation studies. In particular, larger samples might have improved the empirical conclusions for the rather disappointing but common case where the measure of size for selection is the actual population size. This empirical exercise is nonetheless useful to the prospective user since a substantial portion of analysis from survey data involves smaller sample sizes where one is most concerned about the robustness of these distributional properties. The findings seem to suggest that assumed normality may be less reasonable here.

While the work presented in this paper is essential to enable us to apply the full range of classical methods of statistical analysis to finite population samples, there are some practical barriers to a broad acceptance of an approach of the type considered here. One has to do with a question I have as to just how widely the analytic findings extend to settings commonly found in practice. For example, do these distributional properties hold for more complex multi-stage designs when elementary sample sizes are great but the number of sample PSUs is small? Moreover, how are these properties affected when other selection strategies are used to improve the quality of overall estimates or to insure adequate representation of key population subgroups? I am not sure that we know answers to these questions on the basis of the present work, and so clearly more is needed to shed light on these and other such issues. Another practical barrier arises from computational considerations. Finding a suitable approximation to the needed higher-order selection probabilities for Sampford`s method is a step in right direction, but more must be done to address the computational issue of mass-producing estimates and associated variances from permuted sets of data values when application is to a full range of designs with varying sample sizes.

The rest of the papers in the session further illustrate some important design dilemmas but with application to existing samples and the designs through which they are selected. The papers presented by Botman and Bienias each deal with the issue of whether the cost advantage of utilizing sampling frames or actual samples for more than one survey outweighs the design flexibility one loses relative to completely dedicated samples where a design is prepared and implemented for each survey. In the Botman paper the issue is whether to select one sample as a subset of an existing sample, while the Alexander paper considers the matter of whether frames can be effectively shared to generate multiple and largely nonoverlapping samples.

The use of samples for multiple applications has long been a topic of discussion among survey researchers. While the time and cost savings of not having to reconstruct a frame for each survey is a clear and important advantage of these multi-study designs, there are important limitations as well, many of which came out in these two papers. One is that the frame and therefore the sample outgrows its usefulness in time, to the point that after a significant time interval the frame may not only be inefficient for use (e.g., due to decay in the utility of size measures) but it may eventually lead to samples which exclude important segments of the population (e.g., movers or those living where new construction has occurred). Then there is the matter of deciding how to select the second and subsequent samples. This issue becomes a key point in the Bienias paper because of concerns about confidentiality in picking households in more than one sample. One suggestion made is to circumvent the issue by dedicating nonoverlapping subsets of the SSU frame to each study that would draw from the frame, in effect adding a stage for this dedicatory process to the designs. While increased variance is cited as a major drawback to this solution and reasons given for this assertion (some of which I did not fully understand), there are some other statistical considerations linked to this approach that might have been mentioned. For example, how large should individual SSU subuniverse sizes be, since within-PSU sample size needs may differ among surveys (e.g., due to differing stage allocations, frequency of the survey, etc.)? Finally, the design as conceived may not be the most ideally suited for some of its uses. This issue came out in the Botman paper by the fact that the design of the NHIS sample could not provide adequate numbers of black females, aged 15-44 for the NSFG IV sample. Allocation of the sample among stages for NSFG IV was also somewhat constrained by the sample allocation to NHIS. While surveys themselves are multi-purpose in nature and optimum allocations may differ somewhat from measure to measure, there is also the fact that the topic area covered by a survey may cause the most realistic sample allocation for one study to be quite unrealistic for another.

The remaining two papers, one presented by Batcher on an assessment of the taxpayer information service provided by IRS and the other by Hinkins on the design for a study of corporate tax returns, present interesting illustrations of other common design issues brought about by studies that are part of ongoing efforts to gather certain types of data. One issue that both papers address arises from the fact that such studies may be used both cross-sectionally to produce point-in-time estimates and longitudinally to assess temporal trends. In the taxpayer assistance study several comparisons were intended: (1) a given year versus an earlier benchmark period, (2) certain strategically placed time points during the tax preparation period during a given year, and (3) from week to week during any given year. The study of corporate tax returns, on the other hand, is intended to assess trends on a year to year basis. In both studies the importance of designing in overlap in successive samples is emphasized and some of the associated problems of doing so are noted. For example, while in the design for the taxpayer study it is not possible to control overlap among selected inquiries from planted questioners, one can select and use the same set of planted questions through time. This feature does not help with the inference to the specific population of taxpayer inquiries for a given time period, but it does facilitate the comparability of measured quality of tax preparation advice over time. The study of corporate tax returns, on the other hand, has designed into it a system for maintaining a degree of sample overlap over time based on the rates of selection. I wondered about one's ability to control the percentage of overlap under this scheme and whether the commonly used approach of randomly designating a portion of each sample for the subsequent sample and then supplementing the sampling strata for a given period would have provided that control over the amount of overlap in individual strata. To deal with movement among strata from year to year, strata might be collapsed to reduce the amount of between-stratum movement.

Another interesting issue arising in ongoing studies of this type is the matter of how to reconcile its surveillance feature over time with its diagnostic function. This issue is especially apparent in the taxpayer study, where the choice must be made between keeping the same questions to improve comparisons and taking corrective steps with the assisters when weaknesses in dealing with specific questions are pinpointed, thereby invalidating their further use. For example, if it is found that most assisters are misinforming the public on a particular question, does one provide the assisters with a clarifying memorandum to handle that issue, or does one allow the problem to continue in the interest of question comparability over time?

The taxpayer assistance study has the added unique problem of requiring that a contrived sample of questions be chosen from the population of actual past questions, rather than to pick a random sample from the actual set of inquiries made by taxpayers during the period of study. I presumed that it was not possible to do the latter since monitoring or tape-recording would have been required for selected inquiries, steps which could have created both measurement and confidentiality problems. While a sampling of actual inquiries would have been a more direct approach to inference when the goal is to assess how well taxpayers' questions are being answered during a particular period of time, a seemingly reasonable compromise strategy was employed, although a little more information on some key design features would have been helpful. For example, how was the timing for individual test calls determined? Was a single question asked in each of these test calls, how were questions assigned to assisters, and how representative were the selected test questions of all questions in each category? As it were, the overall measure of the quality of taxpayer advice was made by gauging the quality of response to a categorized set of contrived questions and then producing the overall measure of quality by weighting the category-specific assessments by the proportion of actual inquiries in those categories.

In conclusion, then, we see illustrated in these six papers the myriad of issues one faces in designing sample surveys. Moreover, we see that resolution to these issues often requires finding some acceptable middle ground or tradeoff. The papers and this session confirm that survey designs must involve both adherence to principle and the continued search for acceptable compromise. On behalf of the audience, I wish to acknowledge the many new insights the authors have provided to these problems.