

Excerpt from
DISCUSSION

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

Editor's Note: The following excerpt from Kalsbeek's discussion pertains to the Hinkins-Scheuren paper (pages 145-149 of this volume) and the Batcher-Scheuren paper (pages 203-209).

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.