DISCUSSION

Daniel Kasprzyk, Social Security Administration

The Federal administrative record system is a large, complex, and invaluable resource. Although, in general, the records maintained in the system are a by-product of administrative or regulatory procedures, in fact, examples abound in the use of these records for statistical purposes. A recent report [1] sponsored by the Office of Federal Statistical Policy and Standards in the Department of Commerce, identifies major administrative record files, potential uses of these files for data linkages, and technical problems and legal issues associated with their use. Although the use of administrative records for statistical purposes is not a new concept, it has acquired renewed importance in recent years primarily because of greater limitations on obtaining data directly. Respondent burden and limited budgets are serious considerations. Respondent burden can be reduced by supplementing survey reported items with administrative record data for items which are either difficult to obtain in surveys or just as easily obtained from administrative records; respondent burden can also be reduced by the merging of several administrative record systems by means of common identifiers. Administrative record systems also reduce costs by increasing sampling efficiency for certain subpopulations. The importance of the Federal administrative record system to statisticians can be observed at this year's meeting of the American Statistical Association, where at least twenty-five papers are being presented on the statistical use of Federal administrative records. One paper even provides a research agenda for an administrative record population census [2].

Declining funding levels in the Federal budget for the production or maintenance of ad hoc data sources make it mandatory that every effort be made to maintain or improve the quality of administrative record systems. Reduced budgets, however, usually imply that fewer resources are available to research methodological improvements, and to create new data sources--all effort being given to maintaining the status quo. With this in mind, congratulations should be extended to the Internal Revenue Service's Statistics of Income and Research Divisions for their efforts in conducting methodological research to improve and make more readily available data from the IRS Statistics of Income system, a large sample data base of tax return information.

The papers presented at this session, although linked by the common goal of data base improvement, are somewhat different. The Bahnke-Wheeler and Schwartz papers describe a global or systems perspective to the development of Statistics of Income data bases; Harte and Hinkins discuss specific estimation issues in the Statistics of Income program; and Spruill discusses the difficult problem of identifying appropriate procedures to assure the confidentiality of released business microdata files and tabulations. Some specific comments on each paper are given below.

Bahnke-Wheeler

The Bahnke-Wheeler paper describes part of the Statistics of Income processing system for corporations. The paper identifies the complexity of the operation, from the sample selection and sample control process through the consistency edits and review of the output. It also compares and contrasts the most recent developments in the Statistics of Income data testing program. The authors are to be commended for their lucid description of this complex process; enhancements to the system, such as identifying automatic test corrections and providing before/after edit printouts would indeed improve the final product. Yet to understand fully the total system, it is necessary to have some notion of the resources needed to complete the different processing stages, particularly processing time. It would have been very instructive had the authors provided such information as well as addressed other salient issues concerning the processing system. For example: (1) can the problem of identifying appropriate tolerance levels for various processing steps be routinely studied prior to production using either previous year's returns or a subsample of the previous or current year returns; (2) to use current data more quickly, are there times during the processing when an interim product, although incomplete and subject to error, can be delivered to Statistics of Income users?

A second set of omissions concerns nonresponse and editing. What is the magnitude of nonresponse, both item and whole unit, and what is the extent to which manual corrections, consistency edits and imputations are applied to the data? If extensive imputation is planned, it is important to identify imputed values because of the potential effect on sampling error calculations. Unfortunately, the authors provide no indication of the extent to which imputations are flagged.

Schwartz

The Schwartz paper identifies well the goal of an integrated Statistics of Income quality control effort and identifies a number of points which ought to be addressed in their quality control effort. For example, he has raised questions such as: (1) what end products do users need and at what level of reliability, (2) where are the most serious data quality problems in the processing system, and (3) how can resources be most efficiently used to achieve the end product?

As an outsider to the Internal Revenue Service, I am somewhat surprised that the issue of quality levels for various processing phases has not been systematically addressed. Additionally, the lack of use of the source document at various data correction or adjustment stages is also surprising, but no doubt occurs because of operational considerations. Yet I am encouraged by several operational activities. In my experience with large government surveys, I have observed that training personnel for clerical operations has not been given sufficient attention; therefore, it pleases me greatly that efforts to review training procedures and clarify specifications/ instructions are underway. Finally it should be noted that: (1) each item in a document need not have the same weight or importance when defining what constitutes a defective document; perhaps some entries are much more critical than others. This ought to be reflected in the development of quality control procedures; (2) in the past, periodic attempts to measure the quality of final products have been made; it appears that a continuous ongoing project is warranted; and (3) measuring the quality of the file at various processing phases has a utility of its own, but the user is ultimately interested in the final product; adequate measurement of the final product for reliability and validity is his primary concern; consequently the measurement of nonsampling error at this level should continue to be a high priority. [3]

Harte

The Harte paper describes the use of post-stratification strategies to improve the efficiency of the Statistics of Income corporate sample. Unfortunately, the version of the paper which I read prior to this session bears little resemblance to the paper which was presented. I am, therefore, confining my remarks to the version of the paper submitted to me in advance of the meeting. The earlier draft described the results of a Monte Carlo study and compared coefficients of variation for four simulations. Regrettably, the authors never made clear what application the Monte Carlo results might have in addressing the operational problem. Furthermore, a description of the synthetic population of tax returns was never provided; and the relationship between the true Statistics of Income population and the synthetic population was never adequately addressed. In future simulation work, the authors should describe the synthetic populations and how or whether they mirror reality.

Finally, the simulation results presented in the draft paper were clearly not definitive. Much more work and analysis is necessary. The authors themselves suggest that in further work to develop a post-stratification strategy, industry must be augmented with other variables.

Hinkins

The Hinkins paper discusses the problem of developing an imputation strategy to handle the problem of high subgroup nonresponse rates, in spite of low overall rates of nonresponse at the national level. Two imputation procedures are examined for a situation in which total assets for a specific Industrial Division are reported, and the accompanying balance sheet items are missing.

The paper raises several questions: to what extent can the results obtained in the simulation of the Industrial Division corporations which have assets of a specific size be transferred to other Divisions and other size categories? Some transfer may be possible, but only after additional study. Thus, the simulation and analysis should be extended to other divisions and other asset size categories. In creating data sets for the study, different levels of nonresponse should be considered, and the construction and study data sets should be based, to the extent possible, on actual patterns of missing data.

We note that the first imputation method uses the same relative proportion as the previous year. As an alternative, one could model a subsample of corporations which report balance sheet items in the current year and adjust the previous year's proportions accordingly; that is, an element of year-to-year change could be included in the imputation system. The second method should improve distributional analysis because of the inclusion of an error component. Under both methods, however, the traditional hot deck approach adopted for this simulation has the disadvantage of potentially giving rise to multiple use of donors, a feature which leads to a loss of precision for the survey estimators. A review of commonly used imputation procedures and their properties can be found in Kalton and Kasprzyk [4].

It is not obvious from the text whether additional criteria are available for use in this type of nonresponse problem. The simulation and imputation system could be enhanced by the inclusion of other variables which aid in the characterization of the nonresponse problem.

Criteria for measuring the effectiveness of the imputation system were not discussed adequately in the paper; they ought to be. Several, which I offer for consideration: (1) the mean deviation identified as

$d_j = \sum(\hat{y}_{ij} - y_i)/m$, where $\hat{y}_{ij}$ is the

imputed value under the jth procedure and Y_i is the actual value for case i, where i=1,...,m and j=1,...,3; (2) the mean absolute deviation,

$d^*_j = \sum \left| \hat{y}_{ij} - y_i \right| /m$; (3) the

root mean square deviation,

$d'_j = [\sum(\hat{y}_{ij} - y_i)^2/m]^{1/2}$; (4) comparisons

of full sample "true" data estimates of the mean, variance, and covariance with estimates obtained from both true and imputed data, (5) comparisons of distributions from "true" data with distributions based on both true and imputed data, (6) comparisons within the

nonresponse stratum of the true estimates of the mean, variance, and covariance with the estimates obtained solely from the imputed data set, and (7) comparisons of distributions in the nonresponse stratum--that is, comparisons of distributions generated from "true" data with those generated solely from imputed data. Finally, every effort should be made when constructing the simulation data sets to approximate the actual patterns of missing data in the data base, including various nonresponse rates for analytically important subgroups.

Spruill

The Spruill paper is encouraging because it represents a step taken by two Federal agencies, the Small Business Administration and the Internal Revenue Service, to address an issue of substantial importance to researchers both inside and outside the government. Several years ago the Office of Federal Statistical Policy and Standards in the Department of Commerce sponsored work on statistical disclosure and disclosure-avoidance techniques [5]. Since that time, however, not much has been done in this area by the government agencies concerned about inadvertent disclosure. Developing a research agenda for the three different sizes of firms is a sensible approach to the issue of inadvertent disclosure since the problem of publishing the identity of large firms cannot possibly be the same as the problem for medium and small size firms. I am skeptical of the ability to protect the identity of some large firms even with a statistically sound "contamination" strategy. Overlap between variables released and those in publicly available data is another important way to look at the problem; although in reality the overlap surely must be considerable.

My preference for additional work on this subject is to place greater emphasis on the case of variables not normally distributed, based on my guess that many of the types of variables being considered here are not normally distributed. I suspect that the overlap of variables from file to file is extensive; however, it is my belief that the nature of the variables in common is another parameter which should ultimately be considered in the analysis. The duPont Corporation and General Motors Corporation examples are relevant here--the analyst's knowledge of industry and geography is probably sufficient to identify these large corporations successfully with a high probability; if industry and geography were not available, even in contaminated form, more variables would probably be needed to identify the corporations successfully.

An assumption not stated in the paper is whether a one-to-one relationship exists between a publicly available data item and the administrative data item. It is quite likely that definitional differences exist among the various data sets. An additional assumption not stated explicitly is that the types and extent of errors in the measurement of publicly available data are the same as those found in Federal administrative record systems. These are simplifying assumptions not likely always to be true--the actual data problem has an added degree of disclosure protection. Finally, it would have been useful to have been given a more developed discussion of difficulties associated with analyzing contaminated data because most users would have considerable difficulties. Data-related problems confronting analysts become even more severe if contaminated microdata files are released to the public since the pool of potential users would be extensive, comprising a wide range of academic training.

REFERENCES

[1] U.S. Department of Commerce. Office of Federal Statistical Policy and Standards. "Report on Statistical Uses of Administrative Records." Statistical Policy Working Paper 6. 1980.

[2] Alvey, Wendy and Scheuren, Fritz. "Background for an Administrative Record Census." 1982 American Statistical Association Proceedings, Social Statistics Section.

[3] The issues raised in this discussion are general ones and many of the comments could be made about any Federal statistical data collection sysstem.

[4] Kalton, G. and Kasprzyk, D. "Imputing for Missing Survey Responses." 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[5] U.S. Department of Commerce. Office of Federal Statistical Policy and Standards. "Report on Statistical Disclosure and Disclosure - Avoidance Techniques." Statistical Policy Working Paper 2. 1978.

REJOINDER

This reply is in response to the discussion given by Daniel Kasprzyk on five papers dealing with methodological research currently underway in the Internal Revenue Service's Statistics of Income and Research Divisions. The authors of the papers would like to thank Dr. Kasprzyk for his many sound and thoughtful comments. As further clarification on the issues he has raised, we have provided the remarks below.