Prepared for the Internal Revenue Service
U. S. Treasury Department
Washington
16 June 1963

Review of the Sampling Procedures Used by the Internal Revenue Service to Produce Statistics of Income From Individual Tax Returns, with Special Emphasis on Achievement of Quality

W. Edwards Deming, Consultant in Statistical Surveys

I. PURPOSE AND SCOPE

Purpose and scope of this study. This study will deal only with the statistical aspects of the STATISTICS OF INCOME (hereafter S. O. I.) derived from Forms 1040A and 1040, along with various additional schedules such as C and F, for income from business and farm. It does not attempt to cover business-schedules, though it may happen that some of the criticisms and suggestions offered here may apply equally to the processing of business-schedules and to other activities of the Internal Revenue Service (hereafter I. R. S.). The aim and scope of this study as I interpreted the request from the I. R. S. are as follows:

1. To offer suggestions that might lead to improvement of the accuracy of figures in the S. O. I. derived from Forms 1040A and 1040.
   a. to detect the possible existence of biases;
   b. to offer advice to the I. R. S. on possible ways to measure the effects of biases, and on possible ways to diminish them;
   c. to discover ways to decrease the variances of sampling and of small accidental errors of processing.
   d. to review the estimates of these variances.
2. To offer suggestions toward better evaluation of the accuracy of the S. O. I.
3. To seek possible ways to improve the presentation of the results so that prefatory pages of the S. O. I. may better inform consumers concerning the strength and limitations of the figures therein.
4. As a final hope, this study might be of some interest to consumers of the S. O. I., including the economists and committees of economists that work with the I. R. S., devoting their talents to shaping the content of the S. O. I.

The sole reason to undertake this study was to help the I. R. S. to accomplish these aims.

Limitation of scope. This is a statistical study. Its aim is not to tell people that they ought to make more use of the S. O. I. Neither is it to tell the I. R. S. how they might improve the content or classifications of the S. O. I. Economists in government and in business are already familiar with the S. O. I., and are putting them to many uses. The content of the S. O. I., though the responsibility of the I. R. S., as I understand it, is decided mainly on the basis of recommendations from the nation's leading economists, acting as individuals or through the work of committees.

One final word about the scope of this report. It takes the point of view (possibly new) that the consumer of statistical data has a responsibility to inform himself concerning the structural limitations of the S. O. I., as described in the preface thereto, and to possess some familiarity with errors of response, errors of processing, nonsampling errors, sampling errors, and the tricks that fate plays in a complete census as well as in a sample.

The criticisms and suggestions to be offered here fall in line with the nature of large-scale statistical studies. This is not the place to offer untested trigger-happy shots, in the hope that some of them might be worth a thought.

A word on the size of the operation. The first characteristic that impresses anyone who takes a look at the production of the S. O. I. is its size. It can only be described as gigantic, requiring the efforts of 300 man-years per year, in more than 70 locations throughout the country.

Over 61,000,000 Forms 1040 and 1040A are filed annually, and go through various stages of processing. The first step in the production of the S. O. I. is of course the taxpayer's responses on his return, the result of interaction between him and the instructions issued by the I. R. S., explained in some cases by help from an agent of the I. R. S., or from an accountant, or from a friend.

Serialization of returns in many classes takes place upon receipt in 70 locations, followed by selection of the sample. Then comes editing, coding, and tabulation.

The total number of returns processed for the tax-year 1961 for the S. O. I. was 718,000, of which 460,000 were 1040 and 1040A.

I may mention in passing that most of the operations of coding, serialization, and grouping of returns into scores of classes are necessary in the ordinary work of the I. R. S., which is primarily the collection of taxes, not statistics. That is, the serialization in classes is about what it would be were there no S. O. I.

A big project need not suffer blemishes from oversight or from lack of personal touch and care. As a matter of fact, big continuing statistical studies, along with other kinds of mass production, offer avenues for improvement in design from year to year, as well as continual improvement in performance through use of modern methods for the control of quality and of supervision.

Growth of the use of sampling. The phenomenal increase in dependence of government and business on current statistical information, and undoubtedly likewise, to some extent at least, our economic growth, have been possible through advances that have been made in the theory, techniques, and public appreciation of sampling; equally, on better understanding on the part of consumers of data concerning the nature of statistical data.

Even as late as 1940, one hardly dared use the word sample in government statistics. It was advisable, instead, to speak of a cross-section, or of an investigation, or simply of a study. There had of course been scattered examples of probability sampling, through the WPA Census of Unemployment, and in various fragmentary studies here and abroad.

Government statistical series are now indispensable to our way of life, examples being the Monthly Report on the Labor Force, statistics on vacancy, characteristics of the population, payrolls, the cost of living, retail sales, current census of manufactures, and many others, an important one being the S. O. I. Private business spends vast amounts of money annually on single-time and continuing studies of the demands of consumers, and on the performance of product.

Everyone today knows of the powerful impact of the statistical control of quality on the precision, dependability, and economy of manufactured product. Twenty years ago, this use of sampling and of other statistical techniques in industry were in their infancy.

What is sampling? Sampling is the use of statistical theory (a branch of the theory of probability) directed toward improvement of empirical investigation. Specifically, this means more effective use of skills and machines, through improved allocation of effort, and more meaningful interpretation of results. In the hands of a competent theorist, sampling is a tool for efficient administration and management of research.

Sample design, in modern statistical practice, enables one to strike an economic balance between the demand for accuracy and the cost of production. This is so because the statistician may govern pretty accurately in advance, by use of theory, the margin of uncertainty to be expected from sampling, along with the uncertainty that may arise from small accidental errors of a cancelling nature. The same theory enables us to calculate this type of uncertainty from the results of the sample itself, after the returns are in and tabulated, provided: (a) there was reasonable conformance to the sampling procedure as specified; (b) the distribution of the estimates in any cell under consideration is reasonably well understood (which usually means that the cell be not too small; page 27).

Why not 100 per cent tabulations? A common incorrect assumption is that modern computing machines, once in full operation, will render sampling unnecessary, as vast quantities of information may be stored and later recovered in any conceivable combination at the push of a button. This conclusion falls with the major premise. Experience usually shows that the information required today was never collected in the first place, or if collected, was not punched into the card. Moreover, information in the card may not be of sufficient accuracy because of errors in response, or because of errors and gaps in the original records.

A further incorrect supposition is that all uncertainties, even structural deficiencies, along with errors and gaps in response, and the bias of nonresponse, errors in editing, coding, and processing, and everything else that is undesirable will all disappear as large computing machines take over the work. Unfortunately, however, the inherent accuracy in original responses or records, as edited and coded, is the limitation to the accuracy that a machine can turn out.

Probably no set of original records possesses the inherent accuracy and completeness of the tax returns sent in by 61,600,000 taxpayers. There are nevertheless, in these returns, many errors, omissions, and

inconsistencies that if uncorrected would greatly distort many cells in the S. O. I.

Correction of samples drawn from complete files of original records offers a solution to improvement in accuracy. The files or tapes will usually provide a suitable frame for the selection of a sample of the original records (tax-returns, in this instance), along with information helpful for stratification and possibly also for marginal totals to use as a base in the calculation of ratio-estimates.

Such is the case with the S. O. I. In the first place, the tapes used in the I. R. S. contain only information that is requisite for revenue-processing: they do not contain all the information that is necessary for the S. O. I. Moreover, even though the returns subjected to sampling have passed through the normal operation of mathematical verification, further editing is necessary to provide the accuracy requisite for the S. O. I.

Another point is that machine-time, where people keep records of costs, turns out to be expensive. Even where the complete information on a tape is accurate enough to be usable, it is often advisable to carry out tabulations on the basis of a sample drawn from the tape, to conserve machine-time for work that is more productive than mass tabulation. Our own Census tabulation program is a good example. There is a record for every person, and detailed information for 1 family in 4, all placed on the tapes subsequent to final editing; yet, in the interest of economy and speed, and to augment the tabulation program, a significant portion of the tables are produced by sampling the tapes.

II. THE FRAME FOR THE S. O. I.

The universe and the frame for the S. O. I. The frame for the S. O. I. is taxpayers' returns, after they have passed through the operation called mathematical verification, which is, in a word, verification of the taxpayer's arithmetic. The results of subsequent auditing for revenue purposes are not reflected in the S. O. I. (page 33).

The frame is almost a complete coverage of the universe of taxpayers. An exception is a relatively small number of stragglers that come in too late for admission to the S. O. I. for a fixed year. Inclusion of a sample of the stragglers from the preceding year may pretty well compensates for the loss.

Use of the name-file for large returns, described in the preface to the S. O. I., is an illustration of the ingenious efforts made in the administration of the S. O. I. to achieve completeness of coverage. The name-file is a list of names that showed adjusted gross income (hereafter A. G. I.) of $150,000 or over the year before. Any name not found in the sample this year, but which had A. G. I. of $150,000 the year before, calls for a report from the district office. Every effort is made to trace these returns, even if it requires a visit to a district office to recall a return from audit, and to make a photocopy or abstract thereof for the S. O. I.

The stratification and allocation. Briefly, the procedure of selection for Forms 1040 and 1040A consists first of stratification by A. G. I., with breaks at $10,000, $30,000 (new this year), $50,000, and $100,000 (new this year), and formation of a large number of other types of strata within classes of A. G. I., by separation of 1040A from 1040, by presence or absence of Schedule C, or of Schedule F, refund claimed, tax paid in full, part paid, no money received, and of course by the 62 district offices, giving altogether strong economic, demographic, and geographic stratification. These strata would all be required for administrative purposes in the collection of revenue, whether there were S. O. I. or not.

Returns of $100,000 A. G. I. or over are all in the sample,* that is, the probability of selection of these returns is unity. Those between $50,000 and $100,000 are selected with probability of 3 in 10. The probability of selection decreases progressively as the A. G. I. decreases, and is heavier for returns with Schedule C than for returns without it.

This report need not go into a detailed description of the sampling procedures. They appear in the preface of the S. O. I., and in instructions to the district offices and service centers. There would be no point in reproducing them here.

The stratification by A. G. I. is adequate for samples for general purposes. Further breaks in A. G. I. would yield but little additional precision, at the cost of heavy additional administrative loads in the district offices. Further remarks appear on page 20.

* The break for 100% sampling was at $150,000 A. G. I. in years prior to 1962.

Optimum allocation of a sample to strata has meaning only in terms of a stated purpose, and for a given procedure of estimation. The allocation to strata in the S. O. I., under the system of weighting used, I find to be well balanced for a general purpose sample. The four columns that show relative standard errors for the number of returns, for A. G. I., for taxable income, and for tax after credits, appearing in Table V on page 20 in the STATISTICS OF INCOME, 1960, INDIVIDUAL INCOME TAX RETURNS, are remarkably constant over all important brackets of A. G. I. The balance would also be good for other estimates that are highly correlated with A. G. I.

It is of course conceivable that certain specialized calculations that economists might wish to carry out at some time in the future might call for different allocation, if fulfillment of such purposes became over-riding. For example, if one were to work with 1040A returns, with special interest in Pareto curves for certain demographic classes, then there could conceivably be need for a further break in the A. G. I., with different allocation to the strata thus created, in order to give adequate precision to the various parts of the Pareto curves. This is no suggestion that any such thing should be done now. It is only a reminder that stratification and allocation, if optimum for one purpose, may not be so for another. If the purpose changes, the stratification and allocation may change accordingly.

At any rate, an avenue of approach to greater efficiency in the sample-design, possibly more fruitful than the unimaginative suggestion of more breaks in A. G. I., is research to find some figure other than A. G. I. as a mode of stratification--for example, the biggest entry in a return, regardless of what line it appears on (vide the RECOMMENDATIONS, page 25).

The weighting procedure. Estimates are now formed, stratum by stratum, by multiplying the results of the sample in a stratum by a factor equal to the total count of the returns in that stratum, divided by the number of returns drawn into the sample from that stratum (in common terminology, N/n).

Although most of the multiplying factors turn out to be very close to reciprocals of the intended probabilities of selection, discrepancies do turn up here and there between (a) the intended probability of selection, and (b) the ratio of the number in the sample to the total count. It is possible that some of these discrepancies arise from the use of systematic

sampling (page 24). Such a system of selection, applied to incomplete blocks, may lead to important errors in certain types of cells, as incomplete blocks occur mostly in special categories of low frequency, and from returns filed late in the year.

Use of fresh random numbers in any incomplete block would eliminate this source of error, as the allowable margin of departure from the intended sampling fraction could then be calculated from the laws of simple probability. Persistently high or low departure, or a departure beyond allowable limits, would indicate trouble either in counting, or in selection, or in both.

My recommendation to shift to the use of fresh random numbers in incomplete zones has in fact already been largely accomplished at this writing. Further discussion on this point appears under RECOMMENDATIONS (page 24).

Some discrepancies arise from mishaps in carrying out the selection. More of them, I believe, arise from wrong counts of the total number of returns in a class.

The existence of an unresolved discrepancy, whatever be its cause, carries with it the risk of some kind of bias. A fault in selection can lead to almost any conceivable type of bias, depending on what happened. A wrong count, under the present system of weighting, also leads to error.

The instructions and procedures for the S. O. I. call for elaborate and intricate controls and precautions to subject all returns to serialization once and only once. Nevertheless, remarks that appear later on point to the possibility that a very few returns now and then have no chance of selection, or have a double chance (page 23).

Advance data. The I. R. S. has issued for 5 successive years a bulletin entitled ADVANCE DATA FROM INDIVIDUAL INCOME TAX RETURNS, containing skeleton tabulations, based on the regular sample of returns under $150,000 A. G. I. serialized and processed up to mid-November, plus the regular 100 per cent sample of returns $150,000 A. G. I. received up to the 1st December (changed to $100,000 A. G. I. henceforth).

The frame for the advance tabulations of 1961 was about 97% complete. This is unfortunately not the same thing as a sample 97% as big as the sample intended. The remaining 3% of returns, missing at the date of cut-off for advance data, are practically all extraordinary in some way, having failed to pass certain consistency-tests, or being complex and held up for various reasons either by the taxpayer or by the district office. Some are simply late, possibly having come from taxpayers living in far-off countries.

Tabulations of dividends and interest are not shown in the advance tabulations, as they could only be highly unreliable in detailed A. G. I. classes at the date of cutoff.

It requires no great imagination to propose that comparison of successive advance tabulations, cell by cell, year by year, with the final results, might lead to useful laws of extrapolation, as an aid to the consumer of advance tabulations. If useful results were to come from such comparisons, it would be a simple step to advance the advance tabulations, or even to propose two or three successive waves of advance tabulations.

However, to date, there appears to be no ground on which to recommend an earlier cutoff, or waves of advance tabulations.

Ten per cent of the returns fail to pass the consistency tests deemed necessary for satisfactory S. O. I., even though the returns have passed through mathematical verification prior to selection of the sample for the S. O. I. Relaxation of the consistency tests would speed up the advance data. However, having seen the results of consistency tests over many years, I would not recommend relaxation. A better plan, being put into effect at this writing, is to correct a sample of the returns that fail to pass the consistency tests, and to weight the results back into the whole. The sample drawn for correction is as big as can be completed by the date of cutoff for the advance data. (The remainder of those that fail to pass is completed later for the regular S. O. I.)

III. REMINDER ON THE DIFFERENT TYPES OF UNCERTAINTY IN DATA

A word on the genesis of figures in statistical tables. Figures in statistical tables are the end-product of a long series of operations and interactions. The vagaries of fate and chance operate even within the most rigid framework of procedures, however carefully written and controlled.

The vagaries of fate and chance do their work with complete counts as well as with samples. Consider for example, the selection of editors and

coders for a complete census. Certain people answer an advertisement for interviewers. The same way with coders. Some happen to be selected for the work. Others might have been selected. Editors and coders with identical training will now and then have honest differences of opinion on the proper way to handle a problem, as is known by experience. People selected for work produce results that are different from the results that other people would have produced.

A new selection of interviewers and supervisors in a survey will produce a change in results. Even the weather has an effect, whether the survey be a sample, or a complete census. An interviewer finds a particular woman at home merely because a thundershower is in the offing and she decides not to go shopping just now: the replies that she gives to the interviewer will be different in some respects from the replies that would have been obtained from her daughter, who would have given the responses had the thundershower not come up just then. A lawn-sprinkler sends an interviewer around her assigned area in a different direction than she would otherwise have taken, and she finds certain people at home to give responses who otherwise, a few minutes later or a few minutes earlier, would not have been at home. The time of day and a multitude of accidental circumstances affect responses, and the editing and coding thereof. The same types of vagaries affect the processing of the S. O. I.

Even with rules and instructions as full, clear, and rigid as we know how to make them, we find by experience that two people that cover the same area will record different numbers of people resident in the area de facto, and different numbers de jure; and that they will record different figures for their counts of dwelling units, total, occupied, or vacant.

As the size of a statistical study increases, the variance between interviewers and coders may increase because of complexities in supervision: likewise the intraclass correlation from door to door, or (in this case) from one return to another. It is possible, however, to introduce statistical controls as aids to supervision, to hold variances between interviewers and coders to low levels.

Recommendations appear further on toward extension of statistical controls in the production of the S. O. I., with special reference to standards of workmanship in multiple locations.

It is a mistake to suppose that expenses or income from various sources are definitely determinable by consulting records; that anything short of absolute accuracy in a figure can spring only from carelessness, nonresponse, or a wrong entry. Anyone who has ever tried to count the words in a telegram knows that counting things is different from learning the number-system as there must be empirical rules for counting. Is New York one word or two? Does the figure 1063429 count as one word?

It is usually possible, armed with foreknowledge of the nature of the difficulties involved, and by application of statistical methods of control and testing, to produce intelligible forms and questionnaires, and to carry out editing and coding with known degrees of uniformity. A careful job done with intelligent experienced preparation will be different from one turned out carelessly with inexpert preparation. It is a fact, nevertheless, that every figure posted on to a questionnaire, or on a tax-return, is a response to a stimulus. A change in the stimulus (i.e., a change in the question, or even a change in the style of the type, or a change in inflexion of the voice, or alteration of the order of procedure) will bring forth different results.

The same principles apply equally to the most elite physical measurements: the operational definition of any physical property of a material or product is the result of applying specific impulses or tests, and recording what happened.

It follows that there are not absolute figures in empirical data, whether obtained by complete censuses or by samples. The S. O. I. are no exception. This does not mean that data can not be useful. It only means that one must understand the nature of empirical data if he would use them effectively, or if he would offer suggestions or criticisms of methods.

Three types of uncertainty in statistical data. I use the word uncertainty here, rather than error, because not all uncertainty in statistical data, and in the uses made of data, is chargeable to mistakes of man or machine. Much of it is inherent in the structural limitations of a survey, and in the presentation of results. On top of this, the consumer himself may make remarkable contributions to the uncertainty of statistical data, by misinterpreting and misusing them. Intelligent use of statistical data can exist only in an atmosphere of understanding of the various types of uncertainty.

All data, whether obtained by a complete census or by a sample, are subject to various types of uncertainty. This is so, whether the data

come from interviews, questionnaires, or by abstracting figures from original records. The main differences between a sample and a complete count are that (a) the sample has the possibility of being carried out with more care, hence with better conformance to specifications, and with less variance between coders and between punchers than there would be in a complete count, and (b) that the sample is afflicted with a certain amount of sampling error. There are, for our purpose here, three types of uncertainty:

Type I: a. structural limitations in design, content, and technique of interviewing, coupled with failure on the part of the survey-organization to present the data with a clear and full description of their limitations. For example, if one wished to tabulate certain results of the S. O. I. by age of taxpayer in brackets below 62, he would find it to be impossible. Even for ages 62, 65, and 72, the information would not be clean.

b. failure of the consumer to understand the nature of statistical data, and to take into account the limitations of the frame and of the responses or other original sources.

Type II: Identifiable blemishes and blunders made in carrying out the operations of serialization, selection, editing, coding, computation, etc., including eliciting information from the wrong household or record, failure to ask certain questions, or to ask them in the manner prescribed; errors of transcription. These errors have their origin in imperfect workmanship, or departure from specification. There are two kinds of errors of Type II.

a. small errors of a non-cancelling nature (errors that persistently lean in one direction, causing operational bias).

b. large errors. A good example is a single-time blunder, such as copying down the final result of a study as 87.5, when it was actually 85.7. Another example is failure to tabulate some cards, or to tabulate some twice, or to use the wrong weighting factor.

Type III: random variation that arises from (a) differences between the units in the frame that the sample is selected from, and from (b) the inherent uncorrelated or nonpersistent accidental variations of a cancelling nature that arise from interviewers, editors, coders, punchers, and other workers, and from the inherent variability hour to hour of any one worker.

Examples of uncertainty of Type I.

1. Inept specification of requirements. Failure to perceive what information would be useful; publishing (perhaps accurately) information that is of little help to the consumer of the S. O. I.

2. Cutoff date excludes some returns.

3. Undetectable errors of omissions in the taxpayer's return; errors that pass through editing and coding, undetected and uncorrected, or detected and ineptly changed to an incorrect entry.

4. Ineffective rules for coding.

5. Ineffective tabulations, such as classifications and class intervals not well suited to the consumer's needs.

6. Bias that arises from bad curve-fitting; wrong specification of weighting or other adjustment.

7. Failure of the survey-organization to report and clarify the limitations of the figures. The preface to tables should take into account the fact that the users of the figures may lack survey-experience, and may need help to comprehend the possibility of uncertainty in a figure. It should explain what the frame is, what incomes it covers, and the source of the data; whether the returns are audited (they are not) before the S. O. I. are drawn off. It should evaluate and interpret the margin of uncertainty from sampling and from small accidental errors, and the possible effect of blemishes and blunders of a persistent nature that were made in carrying out the processing.

8. Unwarranted deductions on the part of the consumer from failure to read the prefatory notes concerning the content and limitations of the data, and failure to appreciate the nature of statistical data.

9. Failure on the part of the consumer to recognize secular changes that take place in the universe before the results appear.

Examples of uncertainty of Type II.*

10. Mistakes of a noncancelling nature made by the taxpayer.

11. Failure to number all the returns; repeating a whole series of numbers.

12. Failure to subject some batches of serialized returns to the operation of selection; subjecting some batches twice.

---

* The numbers continue, for convenience of reference.

13. Certain other types of mistakes in selection, large, or of a non-cancelling nature.

14. Persistent errors in editing.

15. Persistent errors in coding.

16. Persistent mistakes in calculation and in transcription.

Some remarks about the various uncertainties. Sampling and small uncorrelated or nonpersistent accidental variation (Type III) are, as we see, only one type of uncertainty. This is the type of uncertainty for which a complete body of theory exists, which (a) helps the statistician to design a survey to meet specified requirements, and (b) by which he may in any case evaluate afterward, by mathematical formulas, from the results themselves, the margin of uncertainty from these sources.

The effects of myriads of small accidental errors of a cancelling nature wherever they take place--with the taxpayer, editor, coder, puncher--is included in the standard error. In fact, these small errors of a cancelling nature help to make standard error. They can be held to a minimum by statistical methods of supervision.

In contrast, it is not the function of the standard error to detect persistent omission, or inclusion of material above or below average value, or persistent mistakes in one direction, or an important blunder.* The best way to detect, evaluate, and reduce such biases is to depend on an audit or statistical control (an independent processing of a subsample of the main sample), and on other statistical supervisory tools. Outside sources of information sometimes help.

The insidious thing about uncertainties of Type I, and of the persistent errors of Type II, is their constancy, and the consequent difficulty of detecting them. Tests conducted to demonstrate their absence are oft-times only experimental demonstrations of remarkable ability to repeat the same mistake. To be specific, if the results of a large survey are divided into ten piles at random, or are divided according to the geographic locations of the regions whence they originated, intercomparisons are incapable of detecting a structural defect, because the results in each pile are afflicted equally by the same defect.

---

* There are exceptions. Replicated designs sometimes detect a huge error. One subsample, far out of line with the others, may indicate a huge blunder: see Deming, SAMPLE DESIGN IN BUSINESS RESEARCH (Wiley, 1960), page 72.

Similarly, agreement year after year does not demonstrate the absence of uncertainties of Types I and II.

The distinguishing characteristic of uncertainty of Type I, as already stated, is that it is built into the structure of the study, and that the consumer, for whatever reason, misuses the data. Uncertainty of Type I does not arise from flaws in carrying out the specified survey-procedure: a recanvass (audit or statistical control; vide infra) carried out under the same rules, will not discover a flaw of Type I. STRUCTURAL DEFECTS ARE INDEPENDENT OF THE SIZE OF THE SAMPLE, AND IN FACT INDEPENDENT OF WHETHER WE HAVE A GOOD SAMPLE OR A BAD ONE.

In contrast, a careful recanvass will detect errors of Type II. Responsibility for holding uncertainties of Type II to a minimum rests with the supervision of the job.

This report, being statistical, deals mainly with comments and recommendations concerning uncertainties of Type II and Type III, with only rare suggestions in respect to uncertainties of Type I.

A remark about editing. It is often said that the accuracy of published statistical data can not exceed the accuracy inherent in the source-documents, or in the responses in an interview. This statement is fundamentally true, but it fails to take into account the feats that editors perform, prior to the operation of coding. Editing in large-scale surveys is now usually divided between man and machine. Machines can detect outliers or inconsistencies on a mass scale.

Man and machine possess ability, as editor, under proper rules, to supply entries for certain missing data, and to eliminate some inconsistencies. These emendations, when performed carefully and with competence, definitely produce improvement. The consumer of data owes a heavy debt to the ability of statistical editors.

Yet with all the ability in editing that man and machine display, some errors pass undetected. Moreover, editing is not in every instance an improvement. There is always the possibility that an editor may, in any one instance, supply a figure not as accurate as the one supplanted.

The fundamental statement stands firm, nevertheless, that aside from improvements wrought by editing, the limiting degree of accuracy in the published statistics is the accuracy of the original records or responses.

Specific remarks concerning uncertainty of Type I in the S. O. I.

The figures in the S. O. I. come from unaudited returns. The audit of a return is not finished until the I. R. S. and the taxpayer or his executor or a tax court are satisfied or exhausted. If the I. R. S. were to wait until auditing is finished, the S. O. I. would be ancient history when they appear. The effect of auditing may be important to the consumer of the S. O. I., and in another place I recommend extension of studies in this respect, and that the preface to the S. O. I. explain the main results.

Naturally, the S. O. I. must close its doors to returns that arrive after some specified date. Just as naturally, some returns (in proportion, about 1 in 160) come in after that date, too late for processing in the S. O. I., though every effort is made to include late returns of $150,000, even up to the time of printing. Most late returns are large ones, or are returns from people who have asked for deferment for reasons of health or foreign business that requires them to be out of the country. They are thus presumably different from the returns that make up the S. O. I. In an attempt to offset the loss of the remaining late returns, the S. O. I. for any year contains a sample of the late returns in the previous year.

Certain difficulties in editing and coding may be worth special mention. There are areas of doubt between different kinds of income. Simple wages and salaries probably give but little trouble. In contrast, income derived from a business has vague fringes of doubt. It may be coded as earned by personal services or from a business: an editor or coder may have a hard task to decide.

Confusion between dividends and interest is well known. A taxpayer himself may not know the difference, nor even the editor, fortified with all the rules, instruction, and supervision that the I. R. S. provides.

Interest and capital gain are far apart in character, one might suppose. Yet the distinction between them confuses auditors in the I.R.S., and probably confuses coders even more, and the taxpayer more yet.

Coding the type of business, for income on Schedule C, is difficult, not only in the S. O. I., but anywhere else.

Another point is that the S. O. I. do not include all the personal income in the country: anyone whose income is less than $600 need not file a return.

Moreover, the period of time that the income refers to is not necessarily a calendar year: it may be somebody's fiscal year. This is especially true for returns of corporations; not so much for 1040 and 1040A.

For reasons like those explained in the last section, estimates of dividends and interest derived from two different sources, such as, for example, (a) the S. O. I., and (b) a survey of accounts in banks, trusts, savings and loan companies, etc., could easily differ by $100,000,000, or even in considerable excess of this amount.

Will two samples agree? Will two complete censuses agree? The precision of a sample is not established by comparison against a complete census unless the complete census is the equal complete coverage for this sample. Only in this circumstance will the complete census and the sample have their origin in the same data, definitions, interviewers, coders, and other operations that put a figure on paper or punch it into cards.

An example occurs in the Census, when data that have been obtained for every person, and punched into cards after editing and coding, are sometimes tabulated by means of sampling. The advantage is considerable expansion of the scope of publication, more information, and more information per dollar spent on the Census. This is one of the few instances in which one has the experience of comparing a complete census and a sample, or two samples, that have the same expected value.

One will usually discover that two surveys that appear at first to elicit information in precisely the same way turn out, on closer examination, to be different. The questionnaires will differ in some respects. The surveys will be conducted at different times, carried out by interviewers and processed by editors and coders with different qualifications and with different training and supervision. Small differences in questionnaires, or in hiring, training, and supervision of interviewers and coders will sometimes create big differences in results.

I will go outside the field of income taxes to illustrate the point, how information may differ in the files of the same company. The accounting department of a railway or of a trucking company shows a shipment that weights 120,000 pounds. The original freight bill was one piece of paper. At the end of the line, however, or somewhere along the line, the loads

were diverted to separate points. The files of originating freight bills in the traffic department show one shipment, whereas the files of delivery receipts, and the accounting department, show four shipments.

If a person were to look at the figures for the average weight per shipment, and make no study of the way the records are kept, he might suppose that both figures were wrong. Yet both were correct. It is perfectly natural that figures furnished by the accounting department will disagree to some extent with figures derived from the number of shipments. To interpret either figure, one must understand how it was derived. This is not a fictitious illustration: I drew it from actual experience on this day of writing.

It is always easy to be critical of figures and to point to apparent discrepancies. It is another matter to understand statistics and to use them properly, with due regard to their nature and limitations. The first impulse of a consumer is to look circumspectly at figures, to compare them with related data, or even sometimes to compare them with pre-conceived ideas of what they ought to be--a hazardous proceeding. Comparison with other surveys, when some degree of comparability is justifiable, sometimes helps the survey-organization and the consumer to evaluate and to understand the structure of a survey.

On the other hand, capricious trigger-happy unsupported expressions of doubt about the results of a survey do not improve surveys nor man's ability to understand and use data with discretion.

Comparison of the results of two studies, supposedly giving figures on the same thing, or comparison of the S. O. I. from one year to the next, in any category, requires knowledge of the genesis of the figures.

For example, in my own recent experience, a consumer of data from a sample, writing under the supposition that the average cost of a certain item could only be 50¢, raised a question upon seeing that the average price of this item, as estimated from a study carried out by sampling, was 49.97¢. Investigation showed that the system of charges was not rigid after all; the item was sometimes priced slightly below 50¢. Investigation of the difference thus led to better knowledge about the system for charges, along with better understanding of the results of the survey.

As another example, Business Week for 9 February 1963, page 8, under the heading of "Statistical Confusion" compared two figures on unemployment:

(a) 2.1 million people had been out of work for 6 weeks or longer according to the Bureau of Labor Statistics, while another survey, conducted by the Survey Research Center at the University of Michigan, showed 1.6 million people out of work for 26 weeks or longer. The writer implied that so wide a difference could only indicate statistical inaccuracy. The fact is, however, that the two figures refer to two completely different aspects of unemployment. In spite of the implication, both figures could be accurate, by whatever criterion one wishes to adopt, and they could both be very useful to the expert on problems of the labor force.

Comparison of a complete census and a sample drawn therefrom, or of two samples drawn from the same complete census, is a waste of time if one has for his aim testing the theory of sampling. The fact is, that we know by theory, better than any number of comparisons could ever establish, what the margin of error of a sampling procedure will be for any specified probability, PROVIDED the sampling procedure (selection, weighting, and other operations) as actually carried out, followed specifications reasonably well. The only exception (noted elsewhere) may occur in a cell in which the sample is extremely small, or in which the distribution of contributions is highly skewed, for in such cells the standard error may not be the sole criterion as an indicator of the margin of uncertainty.

On the other hand, it may well be worthwhile to carry out an experiment in sampling in order to learn how to carry out the sampling procedure, and to learn how to do the editing, coding, and other processing, including formation of estimates, and estimates of the standard errors. Comparisons of the sample with the complete count that the sample was drawn from would show the effects of the extra care that is possible in the editing, coding, and other processing of the sample. A study carried out for these purposes will yield much useful information.

IV. RECOMMENDATIONS

General statement. Some of the recommendations that follow here have already appeared on preceding pages. Some of them have in fact already been put into effect during the preparation of this report, or are on their way in. Some of them have been the practice for some time in the I. R. S., but are nevertheless included in an effort to ward off persuasion to possible alternatives. There is no point in bringing up for the sake of argument recommendations that have been presented to the I. R. S. from time to time sounding good on paper, but which in my judgment do not merit discussion.

Recommendations, to be useful, must fall within the bounds of feasibility. The preceding pages are an attempt to lay down terms of reference that recommendations must fit into. Personal visits to a number of district offices and service centers, and numerous conferences with the staff of the I. R. S., along with study of their instructions and plans, have laid further foundation.

Stratification, selection, and estimation. I have examined the theory and the procedure that form the basis for the sampling procedure for the S. O. I., including the stratification, the sampling rates in the various strata, the method of selection, the formation of estimates, and the calculation of estimates of standard errors.

The entire procedure is basically sound, being in conformity with the principles of probability sampling. The number of strata is adequate, especially with the new breaks at $30,000 and $100,000 A. G. I., just instituted this past year. I would not recommend more breaks, under the present requirements. The 62 district offices give strong geographic stratification. Altogether, there are scores of strata based on geography, presence or absence of Schedule C, or of Schedule F, refund claimed, and by other characteristics.

There is a tendency in many statistical organizations, through lack of guidance from theory, to overdo stratification, and to reap only inconsequential gains in precision at considerable cost. Theory may show, in some circumstances, that stratification would be relatively ineffective. On the other hand, theory may show in other circumstances that stratification with proper allocation of the sample to strata is vital, as in the sampling of accounts, business esta-

blishments, farms, and other material in which there is high variance between sampling units, characteristics of income tax being an example.

The finer be the cells in tabulation, the less effective will be the stratification introduced through major categories. This rule does not apply, of course, to cells formed by subdividing a class that was sampled 100%. As another point, the proportion of male and female, the proportion married, the distribution of the number of dependents, and the distribution of income in respect to such characteristics, do not vary much from one category to another. Thus, there is already more stratification of the returns than one would ordinarily specify for sampling, but it comes free of charge, being required for administrative purposes in the collection of revenue, whether there were S. O. I. or not.

The sampling ratios prescribed for the various classes of A. G. I. are well balanced for general purposes, as I remarked earlier (page 7), and I see no need of changing them, although I am in accord with the proposed reduction of the probability of selection from 3 in 1000 to 2 in 1000 for 1040A and for 1040 nonbusiness with total A. G. I. under $10,000.

It is possible, however, as earlier paragraphs suggested, that specialized uses might conceivably in the future call for different strata and allocation thereto. This is only to say that one must be ready to modify any sample-design from time to time to meet changing requirements.

The formulas used in the S. O. I. for the calculation of estimates of standard errors are appropriate. In my opinion, the trifling bias that one might possibly imagine from the use of formulas that are strictly valid under random selection, when the selection is actually patterned systematic, is of no consequence in this application. The estimates of standard errors retain their validity down to cells of small size, even though for very small cells an estimate of the standard error, unaided by other statistical measures, has limited utility as an indicator of the margin of uncertainty from sampling. Possible biases from patterned samples are already being corrected (pages 8 and 24).

In conclusion, my only concern about the sampling procedure is in regard to three points: (1) errors in counts; (2) the weighting procedure, and (3) systematic sampling in every block. Paragraphs ahead cover these points.

Possible errors in counting, and in distortion of weights. The above remarks relate to the design of the sampling procedure. Later paragraphs are directed toward improvement, so far as feasible, of repetitive processes, such as classification, serialization, selection, editing, coding, and punching. There remains the possibility of other types of error in performance, one of which could lie in the counts of the total number of returns in the various classes, plus the risk of double chance of selection, or of no chance at all.

It requires no imagination to offer the suggestion that there may be, here and there, errors in counting the returns in the various classes. One is always safe in suspecting the existence of any kind of error, as almost any conceivable error will make its appearance if we wait patiently. It is another matter to demonstrate the actual existence of an alleged error. It is certainly true, though, as experience shows, that counting by serialization in blocks runs into difficulties unless closely guarded. A skip of serial numbers may go undetected and cause overcount. Duplicate numbers, if they exist, cause undercount. Anyone who has ever worked with serial numbers knows how easy it is for either of these accidents to happen.

When carefully laid out and controlled, counting by serialization is nevertheless about as good a system as man has contrived. Bank clerks count new dollar bills by serial numbers. Officials in charge of the S. O. I. have given careful attention to the serialization, and have installed numerous clever safety devices. Nevertheless, some returns do receive two serial numbers, as necessary routine in the regular work of the I. R. S. It is possible that, in spite of effort, some of these count twice in the total.

It is possible that a batch of returns is now and then subjected twice to the sampling procedure, or not at all. This can happen when a batch of returns, after they receive serial numbers, are for some reason recalled by the district office or by some other section of the I. R. S., and put into a new class, with new serial numbers.

This opens up chances for a number of wrong turns. In the first place, someone may forget to adjust the total count of the category that the returns were removed from. The result would be a wrong count, accompanied by distortion in the weighting of all the cells in the category whence the returns were removed.

The correct procedure would be to sample the returns under the rules of their new class with their new serial numbers, to discard the sample already drawn, and to amend the counts accordingly. Unfortunately, some wrong turns are possible. Regardless of whether the counts are amended correctly or not, it may turn out that both samples get into the stream and are tabulated. Or, someone perceiving that these returns have already been sampled, may decide that one sample is sufficient, and do no further sampling. If someone discards the first sample, we end up with no sample at all from these returns. Though such mishaps are rare, they have been noted.

An example of a rare and inconsequential overcount exists nevertheless when someone, somewhere along the line, beyond the operation of coding for the S. O. I., discovers a flaw in a return, such as no signature. The return is charged out, goes back to the district office, and to the taxpayer. It returns to the district office, and somehow receives a new number. The count is then too big by one unit.

Officials in charge of the S. O. I. have made commendable effort to put into effect and enforce a system to charge out any return once serialized and recalled to a district office, or charged out for any other purpose, and to bring it or a photocopy or an abstract back into the stream for processing for the S. O. I. These efforts appear to be highly successful: few such returns fail to pass through the process of selection.

I may remark, at the risk of saying the obvious, that it is not the function of the standard error to indicate bias from wrong counts, nor from zero or double chance of selection.

I offer on the above points the suggestions that responsibility for serialization, counting, and selection, in every service center, and in every district office, be fixed, so that no question can arise about whom to turn to for information on these operations, nor for investigation and correction. Other suggestions follow.

Fresh random numbers in any incomplete block. Proposal for weighting. A patterned systematic selection (under which the same digits belong to the sample in block after block), in the presence of incomplete blocks, may increase the variance of the number selected, and may lead to high variances of estimates in subclasses. If all blocks, for example, contained only 8 returns, instead of 100, and if the systematic digit were 13 in each block for an intended sample of 1 in 100, then there would be no sample at all. On the other hand, if all blocks contained 14 returns, then the digit 13 would draw a sample of 1:14, instead of the intended size 1:100. These examples are exaggerations, but they illustrate what happens. The obvious remedy is to use fresh random numbers in any incomplete block. I accordingly make here two recommendations:

1. Use of fresh random numbers for selection in any incomplete block. (This recommendation, like some others, is already being put into effect as rapidly as is feasible, at the time of writing.)

2. Use of 1/P for the weight, where P is the probability of selection.

The simplicity of using 1/P for the weight in a stratum would relieve the administration of the S. O. I. of a heavy technical and administrative load from weights that, under the present system, can never be frozen until the last count of a total has been accepted. Re-runs because of revised counts would be obviated, as the weight of a class would be constantly equal to 1/P. Moreover, weighting by 1/P would free the results from errors that arise from wrong counts. In my experience, I would expect these to be more numerous and more serious than mistakes in selection of the sample.

The recommendation for use of 1/P would not apply to advance tabulations.

Continual comparison of (a) the ratio of returns serialized in the various classes, with (b) the number of returns actually selected for the sample, using fresh random numbers in an incomplete block, would provide a useful control over the selection, and over the count-

ing as well, as any persistent deviation in one direction, no matter how small, would indicate trouble in either the counting or the selection, or both.

There are numerous ways to use fresh random numbers in any incomplete block. One may of course use a table of random numbers, but there are other ways. For example, one could merely add the number of blanks in in any incomplete block to the selection-digit specified, and use the sum modulo 100 in every succeeding block until he encounters a new incomplete block; then derive as before a new selection-digit. Use of a table (or of random numbers on a tape where processing is automatic) might be easier where the proportion is high. People can use random numbers reliably on the job.

Possible alternatives to A. G. I. as a mode of stratification. The selection of the sample is based on A. G. I. Instances have been seen where there is huge income from some source, offset by huge losses, resulting in low A. G. I., small probability of selection, heavy weight, and high variance, in the cells that show types of gains and losses. I recommend that the I. R. S. carry on continuing systematic studies on a small scale to discover the proportions of such returns, and their possible effects on the data and on the standard errors. The preface to the S. O. I. should in due time carry a report of this investigation.

It is simple to suggest criterion for selection other than A. G. I. For example, one might hasten to suggest that selection should be based on the largest entry in a return. This would undoubtedly be an improvement, in principle, but it would just as surely run into insufferable administrative difficulties, in the handling of 61,600,000 returns. I nevertheless recommend that the I. R. S. investigate other possibilities in an attempt to discover if there is any criterion better than the A. G. I. as a feasible mode of stratification. For example, one idea to pursue might be to stratify on the absolute value of the A. G. I., positive or negative.

Presentation of results. As John Tukey remarked once in a private communication, the more we know about the inherent uncertainties in a figure, the more useful it becomes. It is a cardinal rule of science

that one should report all the evidence that could possibly affect the reliability of the results that he presents, so that the reader may form his own independent opinion.

A figure standing by itself conveys no meaning. Where did it come from? What is the system of operations that produced it? For the S. O. I., these are the forms, instructions, interviews, and the taxpayer's understanding thereof, the mathematical verifications, editing, coding, punching, and tabulation. The S. O. I. reflects the taxpayer's understanding of the rules.

The authors of the preface to the S. O. I. have gone to considerable effort to tell the consumer what the content is, with paragraphs on pensions and annuities, dividends, exclusion of sick pay, capital gains and losses, depreciation, depletion, contributions, exemptions, etc. There is a table of standard errors. They have tried to conform to the laws of good presentation.

I recommend, however, that the preface set forth some information on the main effects of operational blemishes that occur in processing. The I. R. S. has carried out numerous studies of this nature. More adequate presentation would be possible, however, after the I. R. S. puts into operation suitable facilities for the statistical control of quality (vide infra), and has more information available on the subject.

As a further recommendation, the preface could well include, I believe, information on the main changes that would occur in the S. O. I. were the sample selected after audit (see page 33). Then, too, it would be helpful to see an evaluation of the effects of misunderstandings on the part of the taxpayer (pages 33 and 34).

Presentation of standard errors. I recommend that tables of standard errors show one standard error, not two, and that the preface contain a brief explanation of the use and interpretation of standard error. The strength of the theory of probability lies in its ability to minimize the net economic effect of risks of wrong interpretation of data. Some uses of the S. O. I. require testing a hypothesis; and it may actually be that, in many such tests, two standard errors is about the right multiple to use for minimum economic loss from the risks of accepting a wrong hypothesis. However, the problem that most consumers of the S. O. I. face is one of

estimation, not testing a hypothesis. What would have been the figure in some cell, or what would have been the year-to-year change, had the sample been 100 per cent, and had the processing been carried out under the same rules and with the same care as was exercised on the sample?

The theory of probability can not provide a direct answer to all questions, but it can provide, for almost all the cells in the tables of the S. O. I., a very useful guide to the allowance to make for the range of uncertainty from sampling and from small accidental errors, for any specified risk of being wrong. The only assumptions necessary are that we know the shape of the distribution of the estimates derived from repetitions of the sampling procedure, and something about the distribution of repeated estimates of the standard error of this estimate (e.g., the degrees of freedom in the estimate). It will suffice in most work, in cells where the number of returns is large (say 50 or more), to assume normal theory in the interpretation of the standard error. Actually, 95 per cent of the cells in the S. O. I. for individual incomes meet this criterion.

It is important to remember that, for many of the cells of lowest frequency, the results come from 100 per cent samples and are not subject to sampling variation at all. In fact, the reason why most cells are small is simply because they relate to high values of A. G. I.

Presentation of elaborate theory with respect to standard errors would probably be out of place in the S. O. I. Nevertheless, it might be possible to put down some simple rules in the preface for normal interpretation, with some indication of the conditions under which simple multiples of the standard error have not their normal interpretation. On the other hand, the existing theory in its simpler aspects is readily available in any good treatise on sampling or on elementary statistical theory.

As an inconsequential remark on the presentation of standard errors, the superior (1) which occurs in several cells of Table V on page 20 and Table W on page 20 of the STATISTICS OF INCOME, 1960, INDIVIDUAL TAX RETURNS, where the sample is 100 per cent, could be replaced more effectively, I believe, by 0, as the sampling error is absolute 0. I fear that one's first impulse on seeing the superior (1) is that the sampling error is too big to be trusted, which is the antipode of the meaning intended.

How many standard errors to present. This is always a difficult decision. Standard errors occupy space in a table. Some readers have no use for them: other readers are vitally concerned with standard errors in cells of special interest. No one can foretell all the uses that consumers will make of the data.

It is a fact, nevertheless, that the consumer in practice almost never cares about exactness in a standard error. He is usually interested only in knowing whether an estimate is highly precise, or subject to wide variation. A tolerance of 20 to 30 per cent in a standard error is for this purpose almost always permissible, or even 50 per cent in rare classes.

It is my recommendation to expand Table W on page 20 (another recommendation that is already in effect, to appear in the STATISTICS OF INCOME, 1962). Also, to consider for some tables in the S. O. I., where Tables V and W are hardly applicable, imitation of the scheme of presentation of standard errors followed in the Census of Manufactures.

Research on standard errors in difficult classes. I recommend that there be a continuing study to find useful measures of the sampling variability of the dollar-amounts in small cells in which the distributions are highly skewed. This study might take the form of repeated samples of various sizes from certain classes, especially selected for skewness and oddity, to permit comparison of normal and other theory with the proportion of estimates that appear to fall beyond multiples of estimates of the standard error. Any such experimentation should of course be planned with the aid of theory that will permit some useful generalization.

Quality control of the processing. This report makes no attempt to deal with matters of management and administration, except where statistical techniques are applicable. The S. O. I., as I remarked earlier, is a project of enormous magnitude. Selection of 460,000 returns from a population of 61,600,000 Forms 1040 and 1040A, in 70 different locations, and their subsequent processing up to the point of tabulation, calls for the most approved lines of organization and designation of responsibility.

The I. R. S. is faced with a huge problem of quality control in the production of the S. O. I. The problems are no different in principle from the problems that a large company faces in the manufacture of inter-

changeable parts, or when a number of factories in different locations produce the same product. Fortunately, there is a wealth of statistical theory and a great deal of experience in industry to indicate the general line of attack. The proper theory might possibly give guidance, for example, on the optimum cost to invest on controls per 1000 returns in the various operations.

The Bureau of Customs tests imported goods for quality and for weight at a number of points, to fix valuation for assessment of duty. These tests, whether carried out in Boston, New Orleans, Norfolk, or New York, must not be too far apart. It is easy to imagine what would happen if there were large differences between laboratories in the Bureau of Customs, so that it would be profitable to import, for example, wool, rayon, or tobacco into New Orleans, rather than into Norfolk, because the tests and weights in New Orleans favor the importer. Statistical methods of inter-laboratory tests help to maintain a measured degree of uniformity between laboratories in various locations.

The I. R. S. would be in an uneasy position if any sizeable proportion of year-to-year changes in some cells could well be attributed to editors and coders, or if apparent differences between areas or classes actually arose from editing and coding.

It would not be the most enlightened administration for the national office to specify that service centers and district offices are expected to meet certain specified levels of quality in respect to the sample-location, editing, coding, or punching (e.g., 3 errors per 100 codes). Experience shows that a section of workers, if permitted an allowance of 10% error, will meet the requirement: they will make just under 10% error. If the allowance be lowered to 2%, they will meet it, possibly at an exhorbitant price.

This type of specification is demoralizing and highly unreliable. If it improves quality at all, the improvement can only be temporary and costly. It will not determine what quality is feasible to aim at. Moreover, any system of inspection that will measure an error-rate reliably could be put to better purposes. The only language that is capable of explaining what quality of work is expected from a location or other source is the language of statistical techniques, such as acceptance sampling and control charts.

I recommend that the I. R. S. extend to the processing of Forms 1040 and 1040A, steps that it has already taken in the statistical control of quality in other types of returns. It is important that such work be placed under the guidance of a competent theorist, and that it be oriented toward standards of production in multiple locations.

What is the statistical control of quality? It is a careful examination of small samples of the main sample or of the main job, the purpose being: (a) to evaluate the accuracy of the statistics produced; (b) to discover where instructions and training need revision; (c) to discover the capability of the process, which answers the question, what quality is it feasible to try to achieve, operation by operation?

Samples of editing and coding from every location, selected according to a proper plan, and tested under uniform rules, would provide data for statistical calculations that would point to spots where the work is significantly out of line. To be above or below average is a law of nature: one can safely predict that about half the error-rates will be above average, and half below. To take action merely because an error-rate is above average is indefensible, and will ruin the morale and the work of any organization. What is really important to know is whether an error-rate at a location is SIGNIFICANTLY high or low, and hence indicative of a local problem.

A center where the work is significantly better than average would serve as a laboratory for discovery of ways to improve quality.

Regardless of organization, line of authority, or autonomy at local service centers and district offices, there is a job of quality control to be carried out on the S. O. I. There is no reason to accept persistently inferior work from any one point. If one service center makes significantly three times as many errors as another service center in (e.g.) coding type of business in Schedule C, or persistently throws twice as many businesses into the category NEC (not elsewhere classified), there is definitely a question to investigate. If the administration of the S. O. I. had methods that would point unerringly to the existence of some type of persistent error at a location, the situation would, I believe, correct itself at once if the supervisor at the location received reports, with interpretation.

Essentially, the problem is one of the continual detection of the existence of special causes of variation, as they occur, with immediate feed-back of information to the source, with the aim of discovering the

cause, and of correcting it if feasible. An important by-product of the statistical control of quality would be a continuous report on the quality of the work being performed in the various operations, in each district and center, and within the I. R. S. as a whole. The central administration, and in fact the whole system, would have a continuous display of quality, and an objective answer to the question, "How are we doing?"

Broad-brush tactics are ineffective and demoralizing. Exhortation, admonition, and pleading memoranda sent out from headquarters are helpful in one way: they declare management's appreciation and desire for quality. They may actually produce bursts of improvement, but they furnish no guidance to help the workers to know how they are doing. Too much exhortation may even have a soporific effect. It is not sufficient to issue instructions, even if perfectly clear ones, and to assume that people carry them out, nor to assume that because some person or group did it right before, they will do it right again. Good intentions are not enough. People need contact, and guidance. Maintenance of quality is the result of directed efforts, and utilization of effective statistical methods of supervision.

At best, compliance with instructions can only mean compliance with what someone believes the intent of the instructions to be. The intent can be realized only within limits, and only by observation, statistical test, revision, retraining, and further observation, in a continuing cycle.

The results of a proper statistical program of quality control would in time show themselves in several ways:

1. Improved output and performance, within meaningful quality-standards; not spotted here and there, but uniformly observable over the whole system, obtained through greater efficiency and lower cost through improved procedures, not by pressure on employees.

2. There would be grounds for establishing sensible and achievable quality-standards, by statistical methods that constantly evaluate the capability of the process at each operation.

3. Enlightened morale and incentive, conscious of an overall quality-program, with meaningful quality-standards that are within reach.

4. Objective evaluation of quality and of the accuracy of entries in the S. O. I.

In turn, the official in charge at a service center or at a district office needs methods of the statistical control of quality to pinpoint sources of trouble in his own organization. It is not enough that the work of a group meet requirements of quality: real improvement necessitates use of techniques that will discover sources of error (i.e., certain people, machines, or procedures), however good the overall quality be. Discovery of a source of error, followed by corrective action, points the way to increased output per man-hour of usable product that has the characteristic of constantly improving quality at lower cost.

An additional step in the statistical control of quality would thus be a continuing program of training and adaptation to local needs in the service centers and in the district offices.

Training in the statistical control of quality can be taught at the local level, at low expense. It is well known that the foundation for the improvement of the quality of manufactured product in Japan, and in this country as well, was series of 8-day intensive courses in techniques, infiltrated with basic theory, with a chance for advanced learning by people so inclined.

Improved supervision would be both cause and effect of the program. Records of error-rates, properly kept on a sample-basis, and interpreted with the aid of simple statistical thinking, would show each employee how he is doing, without guesswork. The supervisor would have objective basis for making suggestions, changes in procedure, and transfers. The possibility of favorite treatment would be greatly reduced.

A word on the administration of a quality-control system. Management of sampling requires knowledge of theory, and skilful administration. It also requires funds. The good reputation of sampling is not an accident; it is the result of advances in statistical theory, to be sure, but equally to careful administration, including statistical controls in supervision. The trend, wherever sampling is done carefully, is toward more effective controls.

Controls cost money. It is penny wise and pound foolish to skimp and be tight-fisted on controls. It is far better to cut the size of the sample, accept slightly bigger sampling errors, and to spend the difference on controls to reduce nonsampling errors. There is a far greater hazard of large and insidious bias from repeated persistent errors or large

blunders than there is from sampling variation and from myriads of small accidental errors of a cancelling nature. In any event, the magnitude of uncertainty from sampling and from small accidental errors is known from calculation of the standard error, and from theory.

As the I. R. S. changes to automatic data processing (A. D. P.), there will be a period of turmoil in which the need for controls will be especially acute.

In these days where skilled workers are in heavy demand, the I. R. S. is in competition with other employers, and must find ways to produce the quality required with the skills available. One might be tempted to recommend a higher grade of clerk for carrying out the classification, serialization, selection, and other processing, but it is a fact that mere raise in grade, without a proper overall system of quality control, might only leave things about where they are.

A possible recommendation might nevertheless be to consider the question of whether the grades for editing and coding are high enough.

Effect of audit. As the preface to the S. O. I. explains, and as these pages have mentioned, the sample is selected before audit. Tabulation of audited figures would give different results, and this would be a preferred procedure, were it not that the S. O. I. would suffer intolerable delay waiting for the audit. There would also be difficult administrative problems, but they could undoubtedly be whipped.

Differences between the S. O. I. as carried out, and what it would be if based on audited returns, are undoubtedly of importance to consumers and should be the subject of continued investigation. Actually, the main effect of auditing, so I understand, is on deductions, not on total income. The service center at Lawrence is studying the full paid returns under $10,000, containing Schedule C, that emanated from the district office in 1960. Tabulation of these returns before and after audit, side by side, will be interesting and informative to the consumer of the S. O. I.

I recommend that the I. R. S. institute further studies to learn the possible effects of audit on the S. O. I. Such studies could well commence with classes of returns where the effect might be greatest.

A report on the audit of returns ought to include the results of an organized system of interviews with taxpayers, conducted to discover the effect of misunderstandings at the source, especially in the use and non-use of Schedules C and F, capital gain, interest, dividends, sick pay, and other complexities. If the I. R. S. is already conducting such interviews, on a scale and design that permits conclusions to be drawn with respect to improvements of forms and instructions, the results would be equally useful to the consumer of the S. O. I., and could well be reported in the preface thereto.

Special problems with misclassification and with huge entries. Misclassification and the appearance of huge and unusual units is to be expected in any statistical experience. Misclassification of a sampling unit can occur, and does, by inadvertence, illegible entry, or other failure of man or machine. Two choices of procedure are open in the case of misclassification: (1) do nothing about the weight of a return misclassified; (2) change the weight by a rule, based on theory, that will minimize some stated risk.

As a basic rule, the weight given to any return in the S. O. I. is the weight of the stratum that the return is allocated to. The prescribed procedure, however, makes two exceptions:

> a. When a return with A. G. I. of $150,000 or more is classified by mistake under $150,000, change its weight to 1.00.
>
> b. When a return falls by mistake two or more strata below where it belongs, give it the weight of the stratum one step below where it would have fallen if properly classified. For example, if a return with A. G. I. between $100,000 and $150,000 is classified by mistake under $10,000, then its weight shall belong to the class of A. G. I. between $50,000 and $100,000.

The basic rule for making no change, combined with exception a, is nearly unbiased, as the name-file and other safeguards provide very nearly a complete list of returns of A. G. I. $150,000 or over. This does not mean, however, that the sampling error is small. The fact is that flagrant misclassification, treated by the basic rule, modified by exception a, where applicable, may lead to large sampling error. That is, the result may be

heavy exaggeration in some cells, or exactly the right result in others, as is possible if just the right number of returns like the one in the sample failed to get in.

Exception b introduces bias in order to dampen serious oscillations of the sampling variability that may arise from flagrant misclassification. It is a step in the direction of minimizing the maximum total error, and is defensible, I believe, so long as the number of such misclassifications remains small. No procedure will remove both bias and variance, but exception b provides a choice between two or more evils, when there is no choice but evil.

My recommendation is to make no change at this time in the procedure of weighting misclassified returns, but to keep records that will provide distributions of returns misclassified, study of which, by the proper theory, will enable one to take closer aim at the goal of minimizing the maximum total error for certain characteristics, or for reaching any other goal specifiable in statistical terms.

A further recommendation is to take steps to reduce the number of misclassifications, even though the number be small as it is.

Another form of the same problem, encountered in the S. O. I., occurs when some entry in a return, such as huge capital gain ($110,000, to name an example) is almost exactly offset by a heavy business loss, with the result that the A. G. I. falls low, say below $10,000, or even negative. The probability of selection for such a return would be 3 in 1000, and the weight 333 or thereabouts. The result could be a shock to cells that show capital gain, or which show loss from one source or another, unless there were something like 332 other returns, not in the sample, like the one that fell in.

No action is taken in the S. O. I. in regard to these returns, and I have none to offer except to study the possibility, mentioned elsewhere, of adopting some criterion other than A. G. I., such as the largest entry in a return, for the criterion of classification.

Instructions to the district offices and service centers. Every employee in a service center or in a district office (or anywhere else) has a right to know what his job is: he must know what is expected of him. His work will suffer if he cannot understand the instructions handed to

him, or if written instructions appear to differ from verbal instruction. Instructions that require the employee to guess between two possible alternative meanings will produce errors. Some of these errors may be of a cancelling nature, augmenting the standard error. Others may be noncancelling, causing bias that will go undetected except as discovered by controls.

Instructions that are hard to understand are demoralizing. The argument that instructions are hardly necessary anyhow, as the people know their jobs, and "anyway, we teach them by holding schools and conferences," is a poor excuse for issuing misleading or difficult instructions; it is rather an argument for issuing no instructions at all.

The instructions issued from Washington for the selection of the sample are clearer than most instructions issued in government surveys and by private corporations. There is nevertheless room for improvement. Specific suggestions appear below, referring to instructions entitled SAMPLING SPECIFICATIONS FOR CALENDAR YEAR 1963, dated January 1963. These suggestions are only illustrative, with no attempt to make a detailed list of criticisms and revisions. For example, on page 32:

> (3) Fill out sample selection sheets with the block numbers in sequence. Selection sheets may be held, but no later than August 31, 1963, in order to account for sample returns in block sequence. Pre-numbering of sample selection sheets is permitted, but must be carefully controlled so that the entire selection sheet can be completed at one time, and so that the blank sheet is not improperly mailed to the Statistics Division prior to its completion.
>
> (4) At all service centers, all designated returns are to be sampled before shipment to the district offices. This includes pre-refund returns, returns urgently needed by a district office, service center rejects, and any other designated returns. This rule applies even though the returns have not yet reached the point of sampling.

It is not clear from Step 3 above where the selection-sheets may be held at: is it at the point where these instructions apply, or at some previous point? One could interpret the instruction to mean that the selection-sheets might not even arrive much before the 31st August. And how can returns be sampled before they have reached the point of sampling?

It is essential to be clear on the responsibility for each step, and to leave no doubt about who is to carry it out. The passive voice without the agent is ambiguous, as in the paragraphs above. Return to the virtues of the plain indicative and imperative moods is a good recommendation in instructions, if not everywhere else as well. For example, the above two paragraphs could advantageously appear as follows:

> (3) Fill out sample-selection sheets, showing block-numbers in sequence. You may hold up a selection-sheet no later than August 31, 1963, in an attempt to complete a sequence of block-numbers. It is permissible for you to pre-number selection-sheets (as you may wish to do for convenience and accuracy), but it is vital that you control the numbers, so that you don't mail in improperly a blank sheet to the Statistics Division prior to its completion.
>
> (4) Sample all returns at a service center before you ship them to a district office. There are no exceptions, not even returns flagged for pre-refund audit, nor returns urgently requested by a district office, service center, rejects, or any others.

Revision along similar lines throughout all instructions issued would, I believe, reduce some of the persistent biases that may have their origin in blemishes in selection and processing.

Other lines in the same set of instructions could have a double meaning. For example, on page 7:

> Returns subject to audit selection are to be statistically sampled prior to audit selection with the exception of pre-refund audit returns. Otherwise, they will arrive too late to be used in the statistical program.

"Statistically sampled" is not clear to me, nor what it is that may arrive too late. Moreover, returns designated in the district office for pre-refund audit are actually no exception: the service center is to sample them and then return them to the district office for pre-refund audit. Perhaps the following would be an improvement:

> Returns designated for audit must first go through the procedure for selection of the sample for statistical purposes: then to audit. Reversal of these steps would be the cause of omissions or of intolerable delay in the statistical program. Returns flagged for return to the district office for pre-refund audit are no exception: you must sample them before you return them.

If my suggested revisions misconstrue the intent of the paragraphs quoted, then revision is even more urgent than I had supposed.

Almost every paragraph in the instructions makes a plea for conformance. For example, on page 8:

> District office failure to comply with this rule may result in serious undersampling of certain types of returns.

This is of course a correct statement, but a plethora of special pleas and admonitions is not effective. This one sounds as if deviation from prescribed classes, and from prescribed rules of selection, is worse than deviation from some of the other rules. It is better, I believe, to issue instructions that can be understood and followed, and to expect conformance: to make it clear that any deviation is nonconformance. It is of course well to explain, as the instructions do, what will be the result of this and that kind of departure.

It is a pleasure to repeat that, in spite of these criticisms, which could go on and on, the instructions for the sample-selection in the S. O. I. are better than one usually encounters. Moreover, more than one person in a district office or in a service center remarked that the instructions for the selection of the sample are clearer than most of the other instructions for the processing of income-tax returns.

Acknowledgment. This has been a joint venture with the I. R. S. Everyone there that I have worked with has rendered every possible aid on this study, and has done so with the greatest willingness at any time, nights, Saturdays, and Sundays, at my convenience, with complete disregard of his own affairs. One man went to Chicago with me on a few hours' notice. Every record and every type of difficulty has been thrown wide open for study. Every person that I worked with in the I. R. S. has been eager to seek every improvement possible. People in the district offices and service centers have been cordial and helpful, eager to find ways to do the work correctly.

I may add that people at service centers and district offices charged with selection of the sample exhibited interest and skill in the selection of the sample for the S. O. I. There was no indication that selection of the sample was an impediment to more important duties. On the contrary, the S. O. I. was regarded everywhere as an important product of the I. R. S.