RE-ENGINEERING OF INDUSTRY STATISTICS: MAINTAINING RELEVANCE IN TRYING TIMES

G. Andrusiak, Director, Industry Division,
Statistics Canada, Jean Talon Bldg., 11th Floor

KEY WORDS: Re-engineering, Industry, Statistics, users

I- Introduction

"Why are we producing principal statistics? Nobody uses them!"

It was this comment by one of the members of the Industry Division Management Team that captured our attention at a Divisional planning meeting in the fall of 1992. Although it was a gross overstatement (since principal statistics are extremely important in producing components of the Canadian System of National Accounts) the statement did have the desired effect of focusing our attention on the crucial aspect of users' needs. During the remainder of the meeting we assessed our program's ability (or inability) to meet certain (and changing) users' requirements. We decided, by the end of that meeting, that we needed to review our priorities and re-focus our program.

Let me try to explain why we felt such a review is needed.

Perhaps first and foremost is the fact that the relative positions and structure of the major Canadian industrial sectors have changed over the past few decades. Not only have there been substantial shifts, as goods producing industries have declined relative to services, but the manufacturing, wholesaling and retailing sectors have and are continuing to adjust to an increasingly competitive domestic and international environment. These changes were occurring over a number of decades but have been accentuated in recent times. The Canada-U.S. Free Trade Agreement, the economic downturn and subsequent recovery which has occurred over the past 2-3 years and the proposed North American Free Trade Agreement have been major factors contributing to adjustment. These factors have changed, not only the structure of the industries involved, but to a large extent the information requirements of and about these industries, as well.

The requirements for statistics on businesses (or for that matter most other sectors of the economy) can be sub-divided into two broad categories. The first is the public policy and program requirement. For businesses this relates to information on the economic performance of the economy as a whole and of the multiplicity of industries that comprise it. This type of information is best represented by the various components of the National Accounts and by our current economic indicators (e.g. manufacturing shipments, retail sales). The second category, and perhaps one where insufficient attention has been directed in the past because its needs are not as well recorded, is the area of information requirements of the business community itself. The business community is becoming increasingly aware of the need for improved statistical intelligence as they cope with adjusting to increased competitiveness in a more global economy.

There are indications that both the public and private sector requirements for information are changing. It is perhaps useful to illustrate the nature of these changes through a number of specific examples. In the context of public sector information needs there is a considerable, and perhaps increasing, dependence on the information available from the National Accounts - (e.g. Gross Domestic Product; Input/Output Tables, Income and Expenditure Accounts) in order to monitor the performance of the economy and to make policy and program adjustments. As a result there continue to be changes in what information is required to produce the accounts. For example, Input/Output tables are now disaggregated to a greater extent in response to users' needs. Another requirement is for more provincial level information (e.g. interprovincial trade) in order to compile provincial G.D.P. These types of changing requirements exert pressures on those organizational that produce the building blocks of the accounts.

Closer public sector scrutiny of industrial performance has also highlighted numerous information needs for various groups of business entities that have previously been less well served by the statistical system. Some examples include waste management (where no industrial classification exits), other environmental concerns and the area of business services, among others. Users seem to be asking more questions about what might be termed the "entry" and "exit" of businesses. How many (firms/businesses/establishments) were created in which industry? How many "died" in which industry? How many moved from being manufacturing establishments (an "exit" from manufacturing) to being wholesaling establishments (an "entry" into wholesaling)? In addition to such migration information there are also demands for other business demographic information including various aspects of firm growth, increasing or decreasing contributions to employment and domestic versus foreign ownership. How does the economic performance of various manufacturing industries

compare as we continue to recover from the recession? What makes some firms more competitive than others? These types of questions are not easily answered at present.

In terms of private uses of data there also appear to be changes in the type of information being requested. For example, there are considerably more requests for wholesale trade data than in previous years. Does this reflect a change from less domestic manufacturing to more wholesaling activity? Although there are few concrete measures that can be used to prove this hypothesis, there is considerable anecdotal evidence indicating that this may be occurring.

A second area where there appears to have been a change in the information requirements of the private sector is in the emphasis placed on commodity data. Businesses, when requesting information, are more and more frequently telling us they do not want "industry" data. Instead there seems to be considerable demand for "commodity" data - and commodity data that is not specific to one particular sector (e.g. manufacturing) but instead traces commodity flows, prices and margins from primary production, through the processing, wholesaling, transportation and retailing sectors, including consideration of both imports and exports. Commodity statistics and commodity balances are areas that may not have been priorities of national statistical agencies (except in case of agricultural products), but there seems to be renewed interest as business users continue to request commodity data.

These apparent changes in the private and public sector demands for business data, provide support for a full and thorough re-examination of users' needs and our ability to meet those needs in order to remain relevant.

It may be useful to examine what has happened historically. In the late 1970's program reductions were carried out in a manner that protected, to the greatest extent possible, the integrity of the National Accounts. As a result geographic and commodity detail was reduced. The business community was upset with this loss of data and in some cases undertook to collect and provide it on their own.

Since the early 1980's there has been considerable streamlining of operational processes and some reduction in statistical outputs. In most of these more recent cases of program reductions, there has been an overwhelming reluctance to reduce programs which impact directly on the external users of information. This approach protects the information needs of external clients and minimizes public criticism to reductions in output. However such an approach may also have some undesirable repercussions on other economic series, such as the National Accounts.

The above point serves to illustrate that there needs to be simultaneous consideration of both public and private sector uses and that a balance needs to be struck to optimize the effectiveness of the overall statistical program.

A consideration of the above factors led us (the Industry Division Management Team) to conclude that we perhaps do not have a complete understanding of our users' data requirements, especially in the context of a changing economic environment, and that our program may not be optimal as a result.

It was therefore decided to undertake an internal review of the statistical program of the Division that would start with the crucial aspect of the re-assessment of both public and private sector users' needs (market research).

II- Scope of the Re-engineering of Industry Statistics (RIS) Project

The Terms of Reference for the Project were written in the fall of 1992. The Re-engineering of Industry Statistics Committee was formed shortly thereafter and an overall workplan was developed.

In establishing the Terms of Reference for the Project, a number of important considerations were advanced by the senior management of the Business and Trade Statistics Field:

- One critical consideration is that of ensuring international comparability of industry statistics. This aspect has, of course, been important historically and is now becoming even more important as businesses are operating in an extended international environment. Ensuring international comparability hinges on the adoption of similar classification systems. In particular, it is for this reason that Canada is trying to harmonize its 1997 Standard Industrial Classification revision with that of the United States, Canada's major trading partner.
- The terms of reference for this project also stipulated that consideration be given to not increasing respondents' burden, but that administrative data should be used wherever feasible. Income tax records already play a major role in Canadian industrial statistics; the recently implemented goods and services tax data could be another important administrative source that needs to be explored in detail. In the longer term business administrative sources may provide a means of obtaining detailed commodity and

process information without increasing response burden. Examples of these sources are electronic data interchange (EDI), point of sale (POS) systems and other systems used for "just-in-time" inventories and "just-in-time" manufacturing.

- A third major consideration was that the Industry Division, along with all other business survey divisions, should use a common source as the frame for all surveys. Statistics Canada has made a very concerted effort to consolidate all independent business surveys frames into one comprehensive "Business Register" (BR). This register is now operational and the number of surveys which use it is increasing. The Retail, Wholesale and Manufacturing Surveys as well as the Survey of Employment, Payrolls and Hours (SEPH) are now all using the Business Register. In order to obtain maximum benefit from this central register function, it is absolutely crucial that any revised Industry Statistics Programme continue to use the Business Register as the central frame.

The Industry Statistics Review, as it was first called, was re-named in its infancy to the "Re-engineering of Industry Statistics" (RIS) project. This name change was made because our objectives are essentially to "Re-engineer" --- as the term is commonly used now - in the broadest sense. The project is not only tackling process re-engineering but is starting right from the beginning -- with an assessment of users' requirements.

The overall RIS project can be subdivided into four broad phases:

- consultation with users for identification of their needs;
- review of the expressed needs and decisions on what types of outputs would best meet those needs (essentially this means deciding on the broad parameters of the program);
- examination of alternatives (processes) that would result in the desired outputs (e.g. surveys versus administrative data; quinquennial surveys versus annual, monthly versus quarterly, etc.);
- development of recommendations and an implementation strategy.

III- Managing the RIS Project

The central organizational unit of this project is the Re-engineering of Industry Statistics Committee (RISC). Membership is drawn from all areas of the Division and includes junior as well as senior staff. Although the project is not a true "Participative Work Design Project" which would involve all staff, the members of the Committee make a continuing effort to seek the advice and ideas of all Divisional staff.

There are two full time staff that act as the secretariat to the RISC and who carry out most of the day-to-day work associated with meeting milestones (writing proposals for contracts, developing standard frameworks, setting meeting agendas etc.). In addition to these two full time members there are various sub-groups that are formed on an "as-required" basis.

There is also a senior level "steering committee" which meets every two months to provide direction and advice.

IV- Work Already Performed

The program of Industry Division covers the following areas; manufacturing, wholesale trade, retail trade, energy, construction, logging and mining. The extreme breadth of economic activity that is encompassed by the divisional program injects a very high level of complexity into this project.

Firstly, the program itself is conducted in fashion whereby each of the seven areas works quite independently. As a result there has been little interaction between the groups and virtually no integration of data historically.

Secondly, each of the seven areas faces a large diversity of users and respondents, often with conflicting requirements.

In some cases these are industry associations (e.g. Retail Council of Canada) that can provide a mechanism for consolidating user input into our statistical program. However, such input will often be incomplete since the associations do not reflect the views of all users, but only members' concerns.

The RISC decided that we would have to prepare ourselves for dialoguing with the data users in all seven areas in such fashion as to be able to consider trade-offs between areas. To do this, we felt that staff in each of the seven areas had to have a better common understanding of the entire divisional program - not only of the specific area in which a person is located.

One of the first stages of work was to prepare a "Business Situation Analysis" for each of these seven program areas.

These Situational Analysis Statements covered the historical background of the program, a profile of the known users and uses of the data, the "Product Performance" of publications, CANSIM and special requests. There was also consideration of the program's strengths, weaknesses, opportunities and threats (the so called "SWOT" analysis).

In preparation for user consultations, short (1 page) program summaries have been written. These summaries provide information on the statistical coverage (e.g. establishments); industry coverage (which SIC's), geographical coverage, data collected, frequency, timeliness and dissemination methods. A summary of the most important issues -- as viewed by the program manager -- was also included. When we reviewed these "issues" we were struck by the number of times certain ones appeared in various programs. For example, timeliness is obviously a major concern for users of our present annual surveys. Lack of geographic and commodity detail is another.

Work on the first phase (User Consultations) has been progressing with the compilation of a user list as one of the first priorities. The user list was compiled using publication subscription lists, previous program evaluations, regional offices lists and subject matter staffs' knowledge of their users.

During the first few months of this project there were several meetings with internal Statistics Canada divisions that use industry data. This dialogue proved to be extremely successful for both producers and users, and a greater appreciation of the analytical impediments and production problems was obtained. Although there is continuing, almost daily bilateral contact between Divisions, operational and short-term matters are the focus. These RIS meetings, on the other hand, opened up new channels of communications with a much broader, longer term focus.

As this report is being written the Division is conducting a series of focus groups with business users across the country.

V- What Remains to Be Done

The RISC still has to develop the framework and plans to conduct similar consultation sessions with our major clients in the Federal Government Policy Departments such as Industry Science and Technology, Energy Mines and Resources, Forestry Canada etc. Some of these consultations are likely to be on a bilateral basis; other meetings may be convened for several departments at one time.

Statistics Canada conducts its program in co-operation with the statistical agencies located in each of the ten provinces and two territories. This aspect puts another dimension on the project. Provinces have been asked to participate in this review to whatever level they want or are able to undertake.

The workplan of this project calls for a finalization of user consultations by the end of August 1993.

At that time, we will begin Phase II, one of the most difficult phases of this project -- the translation of the feedback from users into a framework that will allow us to assess the importance of the information/user and to consider the various trade-offs between programs while taking into account quality, detail, frequency and timeliness.

This phase will allow us to decide, in general terms, on what type and mix of outputs would be optimal, given conflicting needs of different classes of users.

Once the broad parameters of the program are established we will move on to Phase III, a consideration of which types of processes that would provide the type of outputs identified in Phase II. Here, we will be looking at such aspects as administrative data, types of surveys (annual, occasional, monthly etc.).

Phase IV, the final phase, will be the formulation of recommendations as well as the development of an implementation strategy. Since it is envisaged that implementation could take up to five years, the final phase of RIS project would in fact be the initial phase of the Division's five year operational plan.

There will no doubt be considerable work to be done in terms of obtaining additional details on users' needs, establishing changes to processing systems, etc. It is really over the next 5 years that process re-engineering would be undertaken.

There are two important factors that we will have to keep in mind as we go through phases II through IV of this review. We will need to ensure that lines of communication with our users remain open and that they be kept aware of our decisions as they are being made. Secondly we must, throughout this process, remember respondents' concerns and their (in)ability to answer certain types of questions. Here we are likely to fall back on the approach often used to decide on questionnaire content.

This consists of a simple four question approach:

1) Do respondents understand the question?
2) Can they answer it? (Do they have the information available to be able to respond?)
3) Will they answer it? This relates both to the sensitivity of the question as well as the time required to respond.
4) Do they understand why these questions are being asked and the purpose(s) to which the compiled results will be put?

It is very likely that a market analyst in a business organization will indicate that certain information is critical for his/her work, while the "respondent" in the same business entity will be either unable or unwilling to respond to that type of query.

VI- What have we learned so far?

This project is still in its infancy so I can only provide some very preliminary ideas at this point.

- There are public/private data use trade-offs. We need to strike an optimal balance in our statistical program.
- Information needs are changing; we have to try to respond to these changes in order to remain relevant.
- We know our users - but we may not have a good appreciation of how they use data and what decisions they make using our data.
- Opening up a dialogue between users/producers of information produces positive results (e.g. we discovered that many areas in Statistics Canada depend on our manufacturing statistics, to validate their survey results, to serve outside clients with more integrated information, etc.).
- We are not experienced at considering users' needs at a broad level. Our focus has tended to be sector specific (e.g. manufacturing) while users want integrated information.
- We have certain issues that are common to many of our business sector programs; for example the timeliness of our annual series. Rather than tackling these problems independently in each area we may be able to arrive at common solutions.
- Involving staff, senior and junior, provides for an influx of many ideas and new approaches while at the same time drawing upon the experience of senior staff - so that we don't "re-invent the wheel".

- E N D -

DATA QUALITY: A QUEST FOR STANDARD INDICATORS

Mesfin Mirotchie
Statistics Canada, Ottawa, ONT. Canada

KEY WORDS: Data Quality, Indicators, Index, social benefit/cost

INTRODUCTION

The issue of data quality is as old as the statistical information production exercise. Strange as it may seem, however, concerns about data quality keep on resurfacing, and continue to occupy the attention of the scientific community (Bonnen, Loeb, Burgess, Groves and Tortora, Tailon, to name a few). This is mainly because of the importance of data quality as an essential attribute of decision-relevant information. The issue is not likely to go away since, among other things, (a) the concepts and methods underlying data collection, processing, and interpretation are continuously evolving, (b) sampling and non-sampling errors are endemic to the process of generating statistical estimates, and most importantly, (c) a growing demand for decision-relevant information increases the value of accurate information and thus increases the cost of data "errors". Furthermore, the data quality varies, among other things, with (a) the level of technical competence, (b) policy measures to which methodological accomplishments are expected to conform, and (c) budgetary and time limitations.

In the absence of a shared data quality standard, the way it is evaluated and communicated to decision makers lacks empirical coherence; hence, it is subject to various interpretations. Public statistical agencies often evaluate the quality of official estimates from the viewpoint of whether the estimates are "fit for use." The estimates which do not meet this criterion are considered to be of insufficient quality for decision making and subsequently are not published. However, users of the estimates do not know the nature of the quality of the data nor do they know the specific measures that were applied to assess statistical errors and measures taken to ensure quality of the data. If there is some degree of responsiveness in the demand for data quality with respect to a change in the value of decisions, the interpretation of data quality should vary with the value of the decisions, and decision makers should be provided with the indicators of the weaknesses of even robust data. The main purpose of this paper is to suggest a set of indicators and a composite index of data quality to evaluate the quality of published data.

A CONCEPTUAL FRAMEWORK

Timely, accurate, credible, and decision-relevant information is a scarce economic resource and the utility of such information is directly connected to the value of the decisions made with it. It is this scarcity that presents the problem of deciding who should provide information, what type of information should be produced, when it should be provided, at what cost, and how to distribute it among private and public decision makers. If the cost of producing such information exceeds its value to the private sector as a private good, then the public sector will endeavour to assume more of the responsibility of producing and making it available to decision makers without exclusion. The value of such information is inextricably determined by its attributes, by the demand for it, and by the extent that it reduces uncertainty of current and future decisions. Increased information on data quality will reduce the cost of using the data in cases where users discount the value of the data due to uncertainty of its quality. Conversely, no information on data quality may increase the cost of using the data in cases where users wrongly assume the data to be of high quality. Of course, all decision makers do not always employ available information efficiently and subsequently fail to optimize benefits. The extent to which the decision-relevant information is valuable depends largely on two factors: (a) the value of the improvement in the decisions to be made and (b) the applicability of the statistical estimates to the empirical reality facing the decision makers.

Demand for decision-relevant statistical estimates is derived from expected improvements in the value of the decisions made with the data. Decisions are inherently multi-dimensional by nature. The use of data without some knowledge about their quality is likely to exacerbate the uncertainty often associated with the multi-dimensionality of decisions. The multi-dimensionality of decisions is even more accentuated as international markets and institutions become increasingly interdependent. The implication of this interdependence is that decisions will have to be "intelligent," on target,

and consistent. The need to make intelligent decisions increases effective demand for, and with it the value of, decision-relevant information that is accurate, reliable, and timely. Whether data are a measure of social variables such as unemployment or housing or agronomic variables such as crop yield or acreage, they summarize collective choices of a society at a given point in time. The data alone are insufficient for decision making. This is because decision makers have no way of determining the adequacy of the estimates for decisions under consideration unless such measures of data quality are made available to them.

The quality of a particular estimate is assessed relative to the underlying true value being estimated. The true value is however usually an unobservable conceptual construct, and this seemingly makes it difficult to evaluate the quality of an estimate. In reality, it is not necessary to always know the numerical value of the "truth" underlying the estimate. It is sufficient to assert only two things: (a) the existence of some unknown true value whose quantitative and qualitative characteristics are captured by an estimate and (b) as the more identifiable statistical errors are removed, the closer the estimate is expected to be near the underlying true value.

Decisions based on an estimate that deviates sufficiently from the truth in either direction are sub-optimal relative to what can be realized with accurate information and are expected to result in a loss of social welfare. The social benefits lost because of the data errors are damages that could be prevented by an improvement in data quality. Although removing more of the identifiable errors is expected to improve data quality, an important question is whether attaining more and more levels of data quality is socially preferable and economically efficient. Beyond some point, increases in the level of data quality entail a decline in social benefit from an increase in social cost of producing data quality. Resources will be allocated efficiently if the value of the resources committed to improving data quality at the margin is equal to the marginal benefits from the decisions made with the data. That is, it is socially efficient to increase or to reduce data quality in such a way that marginal social benefits (MB_s) to decision makers are just equal to the marginal social cost (MC_s) of producing these benefits. This is a guiding economic principal that establishes a limit on the level of data errors removed. The principal, along with various levels of benefits and costs of controlling data errors, is illustrated in Figure 1. The horizontal axis shows an index of data quality (IDQ). The further the distance away from the origin, the higher the data quality resulting from the removal of errors from the data. The vertical

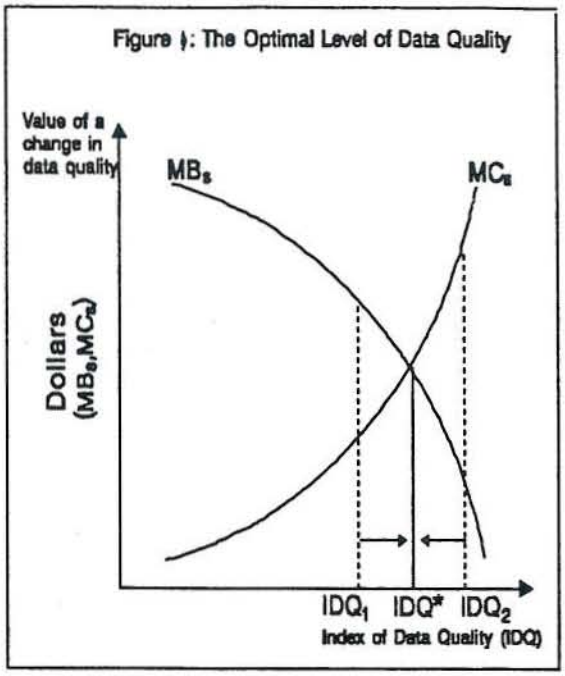


Figure 1: The Optimal Level of Data Quality

axis shows MB_s from, and MC_s of, improving data quality. The optimal level of data quality occurs at IDQ^* where its marginal benefit is equal to its marginal cost. For all the levels of improvements in the data quality up to IDQ^*, say IDQ_1, the marginal social benefit of the decisions made with the improved quality data exceeds the marginal social cost of improving quality of the data. At any point to the right of IDQ^*, say IDQ_2, marginal social cost rises sharply as more and more data errors are removed and it exceeds marginal social benefit. No further adjustment is necessary in the data quality at IDQ^*.

Managing data quality at a level that is socially desirable and economically efficient requires (a) an in-depth understanding of the sources of the data errors and their impact on social wellbeing, (b) an on-going examination of the adequacy of the underlying concepts, definitions and variables selected to represent the empirical universe, and (c) improving the analytical capability of human capital behind the management of data quality and timeliness of decision-relevant information. Decision makers can be made aware of the accuracy of the estimates by way of the indicators of data quality. The next section introduces the idea of developing a standard index of data quality.

JUSTIFICATIONS FOR DEVELOPING AN INDEX OF DATA QUALITY (IDQ)

A great deal of variability exists in the way data quality is evaluated and communicated to the user. This makes it difficult to measure the extent (or degree) of progress, or lack of it, in improving data quality. What is preferred is an agreement on some standard set of indicators with which quality of data can be evaluated and communicated to the user. Among others, two advantages (or reasons) justify the need for an IDQ.

First, the advantages of a composite measure to assess quality of the published data. A composite measure of data quality offers decision makers a clear indication whether or not errors in the data will lead to inconsistent and risky decisions. It would do this in two ways: (a) by providing a concise measure of data quality on the basis of quantifiable sources of errors, and (b) by conveying the level of confidence with which published estimates can be used as input to decision making. In this role, an IDQ enables decision makers to formulate informed expectations about the outcome of their decisions and to evaluate their potential demand for decision-relevant data. **An estimate with a high probability of error, whose quality is not communicated, will be used too often if the decision makers presume the estimate is of a higher quality than it actually is.** The social cost of erroneous decisions may be larger than the benefits.

The indicators of data quality are the result of what has been learned by analysts about real and potential error sources from each prior methodological iteration. The exercise requires re-examination of the estimation process, underlying concepts, methods, and internal checks and balances used for minimizing errors in the estimates. As such, the exercise incorporates an in-depth analysis of sampling and non-sampling errors and interpretation of the accuracy of the data, which is rigorous enough to enhance user confidence in the data. In this sense, the provision of data quality indicators takes the statistical agency one step beyond mere reporting on empirical reality measured by the official data to providing meaningful and interpretive information about data quality to decision makers.

Second, the advantages of standardizing data quality indicators. The ultimate goal of gathering and reporting statistical information is to help the society reach informed economic decisions. Standardization of an IDQ and its component indicators would establish a framework (a) to continuously monitor changes in data quality and (b) to aid internal management of issues of data quality. In the absence of a standard measure of data quality, the concept of data quality indicators (a) invokes different interpretation at each level of data aggregation, (b) varies with specific data for a specific time and space (or geographic unit), and (c) depends on the decisions to be made. An IDQ, together with its component indicators, is therefore a decision tool designed to standardize the way data quality is evaluated by both users of the data and the statistical agencies that disseminates the data. Since an IDQ is unlikely to be stationery over time, changes in it would approximate changes in data quality. Moreover, the development and implementation of an IDQ and standardization of its component indicators would give the maximum internal coherence in the way data quality is evaluated and monitored over time.

THE NEED FOR A STANDARD DATA QUALITY DEFINITION AND ITS INDICATORS

Formulating an indicator of data quality and the nature of its complexity necessitate an achievement of two things: (a) the development of a workable definition of data quality and (b) standardization of the components from which a unique index of data quality will be computed. An ideal index of data quality should be developed from all variables relevant to data quality. However, this is neither practical nor necessary for reasons of tractability, cost, time limitations, and insignificance of the impacts of some errors on the data quality index. Instead, it is expedient to identify a small set of key variables from which a practical composite measure of data quality can be formulated.

For the purposes of illustration, five indicators are proposed as components of an IDQ. The five indicators are: precision, reliability, non-response, timeliness, and residuals. The indicators are selected subjectively. They may not be any better or worse than the ones that are not considered here, but they provide a framework upon which improvements can be made. Attributes of each indicator are briefly described next. The attributes do not pretend to provide global (or all encompassing) meaning to the indicators. They are limited to what would seem to be important from the decision maker's viewpoint, and they abstract only the aspect of the indicator that can be quantified with minimum complexity.

Precision: "Unbiased" estimates with minimum variance are normally understood to be accurate and precise

measures of the underlying truth. Two measures, namely, coefficient of variation (CV) and mean square error (MSE), are often used to indicate the extent of total error in sample observations. For unbiased estimates, MSE can be converted to a measure of precision similar to CV. That is, CV $\cong$ $(MSE)^{.5}$/mean, where mean is a sample mean and CV is defined as a ratio of a sample standard deviation to a sample mean. It describes the amount of total variation relative to the size of the mean. On the other hand, MSE, measures average deviation from the truth in terms of variance and squared bias. As such it combines the measures of precision and accuracy in equal weights. If not pre-determined by a survey design, either one of the two measures indicates an average deviation of the "unbiased" estimates away from the underlying truth. In doing so, they incorporate both sampling and non-sampling (such as coverage and data capture) errors, endemic to collection and processing sample observations. An estimate is judged to be of a good quality if the size of CV or MSE is "low" or of poor quality if otherwise. How high is a high CV or MSE and how low is a low CV or MSE depends on the level of tolerance towards the impurity of the estimate in view of the value of the decisions to be made with the estimate.

It is useful to note here that standardization of a data quality definition and its indicators is not unique to IDQ. They are corner stones for such composite indexes as a Physical Quality of Life Index and a Consumer Price Index. These indexes are far from being perfect, but they serve well the purposes for which they are developed largely because (a) they have become an accepted norm and (b) their definitions and components from which the indexes are developed have been standardized.

Non-response: Data collection procedures are subject to partial or total non-response error for such reasons as refusal, lack of contact, misunderstanding components of a questionnaire, and/or sensitive nature of questions. To the extent that non-respondents behave differently, a systematic bias would be introduced into the estimate. The relative size of non-response will be reflected on the precision indicator of an estimate if adjustments are not made. That is, the higher the non-response rate, the higher the variance associated with an estimate, and the lower its quality. However, The potential damage of non-response on the quality of an estimate may be controlled by imputation techniques.

The important point to remember is that the effect of a non-response rate on the quality of an estimate should be evaluated along with the measure of precision component of an IDQ. A high non-response rate does not necessarily imply bad data quality when its impact is minimized by improvements on the precision of the estimate through a robust imputation procedure.

Reliability: Data are reliable/credible if the estimates are stable in the short run and the frequency of revisions is minimal. In the long run, however, the truth is likely to be unstable within some confidence limits and the estimate is expected to follow that instability. Revisions are important aspect of data quality since they reflect a change in the magnitudes of preliminary estimates, because of the availability of new information, relative to the estimates published in subsequent periods. The implicit assumption here is that all the new information that necessitates revisions has become available within the revision period(s). While data revisions in the long run are seemingly acceptable, frequent revisions in the short run may diminish reliability of decision-relevant data since they are likely to increase uncertainty of the outcome of the decisions. Needless to say, the decision maker is the ultimate judge and it is his/her perception that essentially determines reliability of the data relative to the frequency of and amount of revisions.

Timeliness: Timeliness is an important component of data quality. Its relevance to data quality depends whether it is evaluated from the data producer's or the user's viewpoint. However, such distinction is not as critical as it may seem, especially, for estimates generated by public statistical agencies for the following reason:- public statistics reporting agencies, such as Statistics Canada, often consult with the users of statistical estimates about how the estimates are generated and when these estimates are disseminated. The consultation process establishes, among other things, a mutual understanding between the two entities about the nature of decision-relevant information, data collection vehicles, reference and release dates. As such, the measure of the timeliness aspect of data quality implicitly combines the management of various stages of data collection, processing, estimation, and ability to deliver the estimates within known release dates. Therefore, it is often in the best interest of the producer to adhere to this tacitly "shared contract" since the deviations are likely to mar user perceptions about the dependability of the producer.

In this paper, the measure of timeliness indicates whether or not the decision-relevant information is delivered on an expected date. Two time nodes, namely, reference and release dates, are important from the decision maker's viewpoint. As the time of information availability (release date relative to reference date) is moved into the future, the value of decisions, for which timely information is needed, declines and so does the quality of the information. Reference date has strong implications for the incidence of non-sampling errors such as memory recall error. Thus, timeliness may be measured by the days between the reference date and the release date. The release date indicates whether or/not the data are available to the decision makers on the advertised day of delivery.

Residuals: A residual is a catch-all category of data analysis that captures imbalances between supply and disposition of an estimate at the end of a reporting cycle. As such, it is relevant for estimates generated within an accounting framework such as supply and disposition of grains. The residual indicates a portion of the estimate that is not directly accounted for. The higher the value of a residual related to the estimate, the less accurate the estimate is expected to be.

Indexing Components of an IDQ

An IDQ, the composite data quality indicator, is computed from individual indicators. In practice, each indicator varies between 0 and 1 and is related to the estimates whose quality is evaluated. From the decision maker's viewpoint, an ideal data quality suggests that each indicator attains the value of 0 (or close to it). The value of 0 can be achieved if, for example, the precision attribute of an estimate is high due to a large sample size and there is little or no bias in the estimate; the depth and frequency of revisions between two periods is none (or very small), non-response rate is none or very low; the estimates are released in a timely fashion without delay; and no residuals exist. On the other hand, the estimates are highly "pathological" if each indicator approaches the value of 1.

Finally, the method of indexing individual indicators of data quality and examples of observed indexes and IDQ are shown in Table 1. Each indicator carries an equal weight of importance in capturing the essence of data quality summarized by an IDQ. Unless there is conceptually defensible method of assigning weights to the indicators other than unity, the weight of unity is as good as, or even better than, other weights assigned arbitrarily to the component indicators. In the Table, examples 1 and 2 can be considered as indicating either the quality of an estimate in two time periods or the quality of two estimates in one time period. In either case, the quality of an estimate represented by IDQ_2 = 12.50 is better than the quality of an estimate indicated by IDQ_1 = 4. The higher the value of an IDQ, the better the quality of an estimate.

CONCLUDING REMARKS

Despite statistical agencies' relentless pursuit of improvements in the quality of official statistical estimates, there still does not exist a consensus on the definition of data quality and its indicators. If substantial progress is to be made in the way data quality is evaluated and communicated to decision makers, the definition of data quality must be standardized; a composite measure of data quality must be developed, and a small set of standard data quality indicators needs to be determined. The discussion in this paper illustrated each of these factors and provided a conceptual economic model for achieving socially efficient level of data quality.

REFERENCES

Bonnen, James T. "Improving information on Agriculture and Rural Life." *American Journal of Agricultural Economics*, 57(December, 1975):753-763.

Burgess, R.D. "An Examination of Statistics Canada's Data Quality Release Criteria," *Proceedings of Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, Ottawa, Ontario, 1991: 273-285.

Groves, M. and Robert D. Tortora, "Developing a System of Indicators for Unmeasured Survey Quality Components." Memo, U.S. Census Bureau, Washington, D.C. 20233, USA, (Year not indicated).

Loebl, Andrew S. "Accuracy and Relevance and the Quality of Data," in Liepins, Gunar E. and V.R.R. Uppuluri, *Data Quality Control: Theory and Pragmatics*, Marcel Dekker, Inc., New York, 1990: 105-143.

Tailon, Jacque, "Monthly Gross Domestic Product by Industry: Quality Assessment." Statistics Canada, Ottawa, ONT. K1A OT6, Cat #15-001 Monthly:128-139.

Table 1: A Hypothetical Illustration of Indexing Data Quality Indicators and Computing an IDQ

Indicator	Index	Example	
		1	2
Precision	Ratio of standard deviation to a mean of an estimate	.20	.05
Non-response	Ratio of partial and/or total non-response to sample size	.25	.15
Reliability	Ratio of estimates revised to total sum of published estimates	.35	.10
Timeliness	Ratio of the number of late days required to release data to total days between contiguous reference and release dates	.15	.05
Residuals	Ratio of the residual amount that is not accounted for directly by an estimate (e.g. feed, waste, dockage) to an estimate. Note that in a supply and disposition analysis, an estimate is the sum of production, change in inventory and imports.	.30	.05
Composite Indicator of Data Quality $IDQ_i = N(\Sigma d_i)^{-1}$ =		**IDQ_1** **4.00**	**IDQ_2** **12.5**

BUSINESS FINANCIAL STATISTICS PROGRAM - A STATISTICAL DATA OUTPUT MODEL

Jack Wilson
Statistics Canada, IOFD, B8-10th Floor, Jean Talon Building, Ottawa, Ontario, K1A 0T6

The Business Financial Statistics Program in Canada covers all incorporated for-profit businesses that operate in Canada. These businesses encompass activities both in the financial and non-financial industries. Statistics Canada has recently developed and implemented new standards for this program. These standards cover three elements of the program, 1) the unit of observation, 2) the industry groups, and 3) the data content or core set of statistics. The standard business unit of observation used to collect and tabulate financial data from businesses is called a "STATISTICAL ENTERPRISE" which will be referred to as an ENTERPRISE[1] in this paper. The "Canadian Standard Industrial Classification for Companies and Enterprise" is used as the basis for the published industry groupings and the industrial classification of enterprises. It should be noted that this is a separate classification from the one used to classify "Establishments" in Industry Production Statistics Program. The subject of this paper is the standard core set of statistics for all industries both financial and non-financial as described by the statistical data output model.

BACKGROUND

The Business Financial Statistics program goes back 40 years. In the early 1950's at the time of its beginning it was restricted to a business profit survey covering selected industries - mainly in manufacturing. Over the years the program changed and grew. In the formative years the program existed only to feed business profit numbers into the Canadian System of National Accounts (CSNA). Therefore the content of the Income Statement was designed solely for the purposes of the National Income and Expenditure Accounts of the CSNA. The non-financial industries surveys went through a major expansion in the early 1960's when the Balance Sheet accounts were added to meet the needs of the newly launched Financial Flow Accounts in the CSNA.

The financial industries surveys were gradually introduced into the program during the 1960's and 70's. Virtually every industry in this sub-sector had a unique set of accounts because of the unique nature of financial services provided by enterprises in these industries and different regulatory reporting requirements. The program was a collection of industry specific surveys where the content, data definitions and concepts of the surveys was independently developed. At the end of this period of history it was apparent that the differences between industries rather than the similarities were emphasized. This made inter industry comparisons difficult if not impossible for a number of important performance indicators. During the past 25 years there was a steady growth in the number of users of these statistics. With this expansion there was an increasing diversity in the uses made of the statistics. Some of the user demands required additions or modifications to the surveys. It came to the point where the increased demands created conflicting requirements that became very difficult to accommodate. These demands required expanding survey questionnaires which increases survey respondent burden. This had become a major issue which was addressed in the most recent program redesign. To reduce survey respondent burden and to focus on the most commonly used elements of the statistics, it was decided that the program content and the number of survey questions had to be reduced. This brings us to the development of a "Statistical Data Output Model".

BUSINESS FINANCIAL STATISTICS USERS AND USES

The Business Financial Statistics program was recently redesigned. At the outset of the redesign project there was a thorough review of the fundamental objectives of the program. We went back to basics including consultations with the major users of the statistics. Common elements and the most widely uses statistics were explored with the constituent user groups. These groups could be put into the following categories.

1. Canadian System of National Accounts
2. Macro economic forecasters
3. Public policy analysts both at the sector level and individual industry level
4. Industry associations
5. Credit granters (lenders to businesses such as Banks)
6. Investment Dealers (investment decisions and capital markets)

Standardization of the data content would facilitate inter industry comparisons and aggregations of industries which was identified as a requirement of most users. This was the main stimulus to develop a standard data output model. Superficially there appeared to be differences in the needs of Economists

and Financial Analysts. But underlying these apparent differences we identified a lot of similarities in the data and concepts used by these two groups.

INTERNATIONAL COMPARISONS

In the mid 1980's when the scope of the program redesign was determined the international aspect of Business Financial Statistics was not considered a major issue. However, we now see international comparisons of the financial health and performance of the private business sector gaining interest. This is due to the recent moves to freer international trade and greater international mobility of capital and production activities of business enterprises. In the future we would like to look at the Business Statistics of other countries, in particular our major trading partners, to asses the feasibility of these comparisons and to promote greater international harmonization of standards used in this statistical program.

ANALYTICAL FRAMEWORK AND CONCEPTS

In developing the analytical framework for the indicators and measurements of financial performance and financial health some of the most commonly used Financial Statement accounts and financial ratios were selected:

FINANCIAL PERFORMANCE

1. Profitability
    a) Operating Profits[2]
    b) Net Profit[3]
    c) Rates of return
    d) Dividend pay out rates
    e) Cash generated (net cash flows) from operations
2. Operating efficiencies and operating leverage
    a) Accounts receivable turnover
    b) Inventory turnover
    c) Operating profit margin (operating leverage)
    d) Operating profit per $ of Assets

FINANCIAL HEALTH AND STRENGTH

1. Liquidity and solvency
    a) Working capital ratio
    b) Debt to equity ratio
2. Capital structure
    a) Debt and equity financing
    b) types of debt financing
    c) positive/negative capital leverage

The most widely used indicator of financial performance is "profit". Investment decisions are made based on a businesses ability to generate profits. Profits and the expectation of profits are essential to attract the financial capital needed to finance economic activity in the private for-profit business sector of the economy. Entrepreneurs and investors must have confidence that their investment in a business will generate adequate levels of profit to yield a return commensurate with the degree of risk. Rates of return are derived by calculating a ratio of profits to owners' capital investment. Profits are also monitored at the industry and sector levels for the purposes of public policy analysis, and economic forecasting. Profits are used as the measurement of business income in the National Income Accounts of the CSNA. There are several measurements of profit that are produced in this program to suit the different applications. The two most widely used are "Operating Profit" and "Net Profit" that are defined in the footnotes. The CSNA Economic Production Accounts uses a profit called "Net Operating Surplus" which is, in a practical sense, close to "Operating Profit" in this program. Net cash flows represent the net cash inflow as a result of the operating activities of a business. This number is calculated by adjusting the net profit which is measured on a accrual basis to profit on a cash basis.

Operating efficiencies and operating leverage relate to the notion of maximizing the accomplishment with a minimum amount of effort. In the context of financial performance efficiency and operating leverage is analyzed by looking at the amount of revenue generated from the sales of goods and services for a given level of expenses. The objective is to maximize revenue with a minimum of expenses. Another way of analyzing this issue is to compare the operating profit to the operating revenue (sales of goods and services). The indicator is called the "Profit Margin" ratio and it is derived by calculating the ratio of operating profit to operating revenue. Operating efficiencies are also looked at from the point of view of asset turnover rates.

Financial health and strength issues are looked at from the point of view of liquidity and solvency. Liquidity analysis focuses on the assets of an enterprise and indicates the portion that are liquid. Liquid assets are cash, short term marketable securities, and other assets that are easily converted to cash and that will in fact be converted to cash in the near future. Solvency tests deal with an enterprises overall financial viability. This analysis looks at the issues of an enterprises ability to meet its financial obligations in the future. What is the relative risk of financial failure? What the chances of an enterprise surviving the troughs in the business cycle or other adverse economic conditions? The most commonly used indicators of solvency both short and long term are the working capital ratio and the debt to

equity ratio.

A core set of statistics was established to provide the data required by this analytical framework.

THE MODEL

The model is a standard set of account classifications and definitions to be used when providing business financial statement statistics. The standard data output model not only facilitates inter industry comparisons it takes into account other financial statistics users that are interested in only one industry, or specific issues that go beyond the core issues identified in this paper. In particular the unique features of financial industries requires the inclusion in the model of major elements which are applicable to financial industries only. One of the challenges was to integrate the unique accounts of specific industries into the common framework.

All enterprises maintain accounts and apply accounting rules which allow them to produce annual "general purpose financial statements". The accounting rules and financial statement elements are based on the codified business accounting standards for all for-profit businesses in Canada. The standards are commonly referred to as "generally accepted accounting principles" (GAAP). The accounts, definitions, and accounting principles in these standards were used as the primary reference in building the accounting structure of the model.

The model includes three financial statements which incorporate the following elements and accounting equations:

BALANCE SHEET

Asset accounts = Liability accounts + Shareholders' Equity accounts

INCOME STATEMENT

1. Operating revenue - Operating expenses = Operating profit
2. Operating profit +/- Other revenue, expenses, gains and losses = Profit before income tax
3. Profit after income tax +/- income/loss from unconsolidated subsidiaries, and extraordinary gains and losses = Net profit

STATEMENT OF CHANGES IN FINANCIAL POSITION

Sources of Cash:

Cash from operating activities
Cash from financial activities
Cash from deposits

Applications of Cash:

Cash applied to investing activities
Cash applied to acquire capital assets
Cash applied to lending activities
Cash applied to pay dividends

Net increase/decrease in cash balance from the beginning to the end of the period.

STANDARD CHART OF ACCOUNTS FOR SURVEY QUESTIONNAIRES

The Business Financial Statistics Output Model is intended to prescribe the standard accounts used for all industries. Some of the published accounts and ratios are calculated from the data collected in the financial statement surveys conducted by Statistics Canada. Other data not included in the model but required by the CSNA has to be included in the financial statement survey questionnaires. So the data collected from enterprises is not identical to the output model accounts. To accommodate these differences a separate document was developed that sets out the list of accounts from which questions could be drawn for survey questionnaires. This document is referred to as the "STANDARD CHART OF ACCOUNTS". An essential element of this development was consultations with representative groups of enterprises of different industries. This was to minimize respondent burden in terms of survey questions that could be easily answered. Both the standard chart of accounts and the data output model have a hierarchical structure and numbering system. The highest level elements and categories are common to all industries but as one moves down the structure one finds that some of the more detailed accounts are unique or significant to certain industries.

NOTES:

1. An enterprise is an economic unit that consists of one or more entities under common ownership and control. It's management is separate and autonomous from more senior levels or its parent corporation in terms of decision making powers. It is empowered to enter into transactions covering investing activities, financing activities, and operating activities. A full set of consolidated financial statements are prepared for this economic unit for outside users including investors and credit granters. For most of the largest enterprises this unit represents a family of corporations under common ownership and control.

2. Operating Profit is a residual. It is the excess of Operating Revenues over Operating Expenses. Operating Profit represents the net results of the operations which takes into account all economic transactions and events (revenues and expenses) associated with the principal and ordinary activities of

an enterprise. For non-financial enterprises secondary or ancillary activities, such as dividend and interest income, transactions of a capital nature (capital gains), and non-recurring extraordinary transactions and events are not included in the measurement of Operating Profit. Interest expense related to debt and borrowing is also excluded from operating expenses in the calculation of Operating Profit.

3. Net Profit represents the residual net earnings from operations and all other revenues, expenses, gains and losses from secondary and ancillary activities that accrue to the owners (shareholders) of the enterprise. Other expenses include corporate income tax, and interest expense on long term debt and loans. Dividends distributed to shareholders are not deducted in the calculation of Net Profit. Conceptually Net Profit is supposed to represent the increase in wealth of an enterprise over a period of time that accrues to it's owners (shareholders). It is illustrated by the increase in Balance Sheet account called retained earnings which is part of the owners equity in the enterprise. This increase is before dividend distributions of earnings are deducted from this account.

STATISTICAL PROCESS CONTROL IN MINERAL INDUSTRY SURVEYS

Ching Yu, Sandra Absalom, Lynne McClaskey, & Jeff Busse, U.S. Bureau of Mines
Ching Yu, 810 Seventh St., Washington, D.C. 20241-0002

OVERVIEW

A major issue related to government surveys is the timely release of survey statistics to the public. Each month the U.S. Bureau of Mines (USBM) conducts 22 Mineral Industry Surveys (MIS) and publishes a report on each for public dissemination. These monthly surveys collect production and consumption data from more than 2,600 establishments engaged in mining, mineral processing, and other mineral-related activities. The process of publishing the MIS data consists of four sequential phases of activity. These phases involve survey closeout, data processing, preparing the report for publication, and distributing the report to the public.

The USBM established targets for completion of each phase of the publication process. The Branch of Statistics and Methods Development within the USBM routinely evaluates the success of each survey in meeting the targets. MIS timeliness reports are prepared and distributed on a monthly basis to appropriate organizational units.

The USBM has recently adopted the concepts of Total Quality Management (TQM). Customer satisfaction, employee involvement, and continuous improvement elements are being emphasized. Systematic approaches to quality improvement using statistical methods are being applied in a variety of USBM activities. The purpose of this paper is to examine the potential for Statistical Process Control (SPC), which is the heart of TQM, to facilitate the MIS publication process.

Historical data from the monthly MIS timeliness reports were used to construct SPC charts for the 22 surveys. The charts were used to identify MIS processes that are under control and those which require special attention to bring into conformance with publication timeliness targets. Subsequently, the SPC charts were used to set priorities and evaluate the success of survey process adjustments.

DATA SOURCE AND VARIABLES USED

Producing an MIS report for publication requires four phases of activity. The activity phases and the timeliness target period for each phase are listed below:

Document closeout:	28 work days
Tables completed:	5 work days
Forwarded for printing:	5 work days
Distributed to public:	7 work days

This study encompassed analysis of historical data on the timeliness of each phase of the 22 monthly MIS. The source of historical data was monthly MIS Timeliness Reports from November 1989 to December 1992.

STRATEGY EMPLOYED

To identify and rank MIS which were experiencing problems in meeting timeliness targets, a statistically based method was developed. The **Mean Workdays of Completion (MWC)** of each activity phase was computed for each of the 22 monthly MIS, as listed in table 1 and plotted in figures 1-4. Statistical methods were used to determine which MIS processes deviated significantly from normal, as follows.

For each activity phase, the grand mean was calculated (table 2) and cases of significant deviation from the norm were identified as those with MWC equal to or exceeding one standard deviation (+1.00 sigma) from the grand mean. The MIS thus identified (table 3) were ranked highest priority for applying SPC techniques to improve timeliness.

A second order ranking was identified as those MIS having MWC which lie between +1.00 sigma and the grand mean. A third order ranking was identified as those MIS having MWC which lie between the grand mean and the timeliness target. The remaining MIS, having MWC equal to or better than the

Table 1.--Mean Workdays of Completion for Each Activity Phase of Monthly MIS Production (November 1989 - June 1992)

Activity Phase / Commodity (Symbol)	Document closeout	Tables completed	Forwarded for printing	Distributed to public
Aluminum (AL)	27	6	5	6
Cement (CM)	25	4	3	7
Chromium (CR)	39	11	7	7
Cobalt (CO)	40	23	6	6
Copper (CU)	35	18	22	7
Gold & Silver (AU)	33	14	8	6
Gypsum (GY)	32	2	3	6
Iron & Steel Scrap (FE)	42	29	6	6
Iron Ore (IO)	21	26	20	6
Lead Industry (PB)	35	7	14	5
Lime (LM)	22	4	2	7
Manganese (MN)	30	5	8	6
Molybdenum (MO)	36	14	8	5
Nickel (NI)	23	5	7	7
Phosphate (PR)	20	2	2	7
Silicon (SI)	35	5	5	6
Sodium (NA)	19	3	2	7
Sulfur (S)	21	3	2	6
Tin (SN)	28	5	9	6
Tungsten (W)	32	8	6	5
Vanadium (V)	29	11	5	6
Zinc (ZN)	32	6	13	6

* numbers are rounded.

Table 2.--Parameters for Ranking MIS*

MIS Activity Phase	+1 Sigma	Grand Mean	Target
Document Closeout	36.53	29.82	28
Tables Completed	17.36	9.60	5
Forwarded for Printing	12.76	7.41	5
Distributed to the Public	6.83	6.18	7

* measured in workdays.

Table 3.--MIS Ranked for Management Attention

MIS Activity Phase	Priority Order 1	Priority Order 2	Priority Order 3
Document Closeout	**Chromium** **Cobalt** **Iron & Steel Scrap**	Copper Gold & Silver Gypsum Lead Industry Manganese Molybdenum Silicon Tungsten Zinc	Vanadium
Tables Completed	**Cobalt** **Copper** **Iron & Steel Scrap** **Iron Ore**	Chromium Gold & Silver Molybdenum Vanadium	Aluminum Lead Industry Tungsten Zinc
Forwarded for Printing	**Copper** **Iron Ore** **Lead Industry** **Zinc**	Gold & Silver Manganese Molybdenum Tin	Chromium Cobalt Iron & Steel Scrap Nickel Tungsten
Distributed to the Public	**	**	**

** all MIS meet the timeliness target, no analysis is needed.

Fig. I Document Closeout

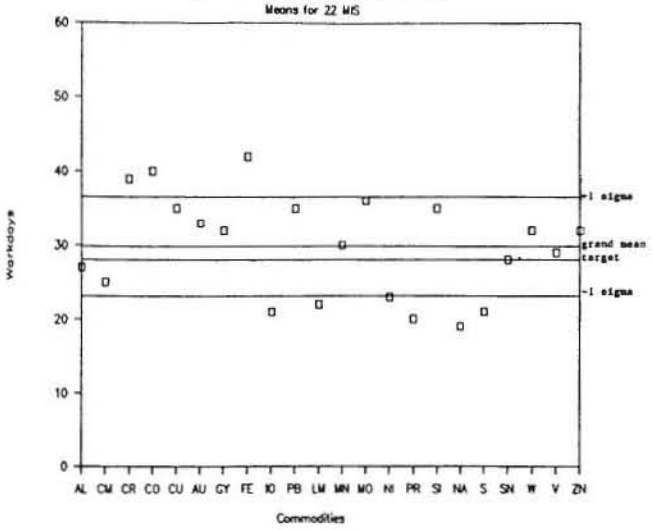


Fig. II Tables Completed

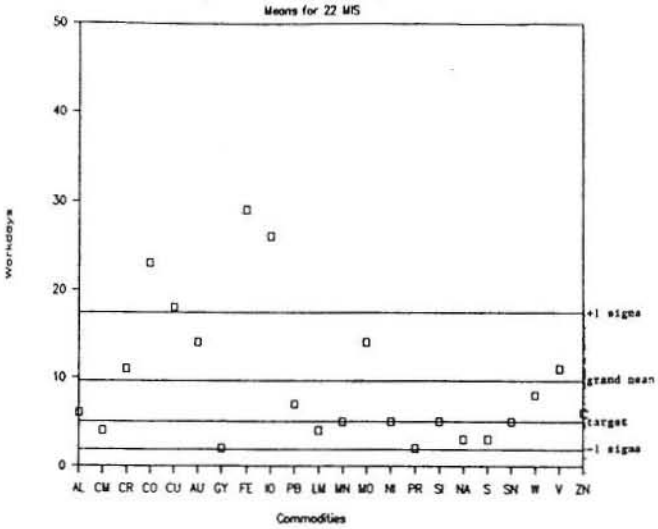


Fig. III Forwarded for Printing

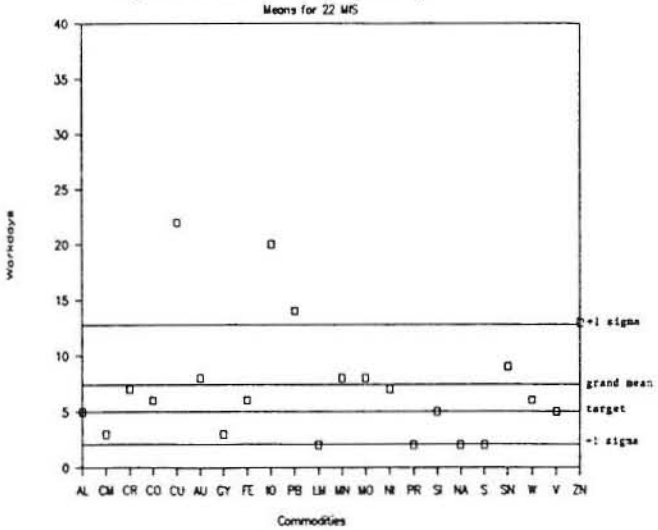
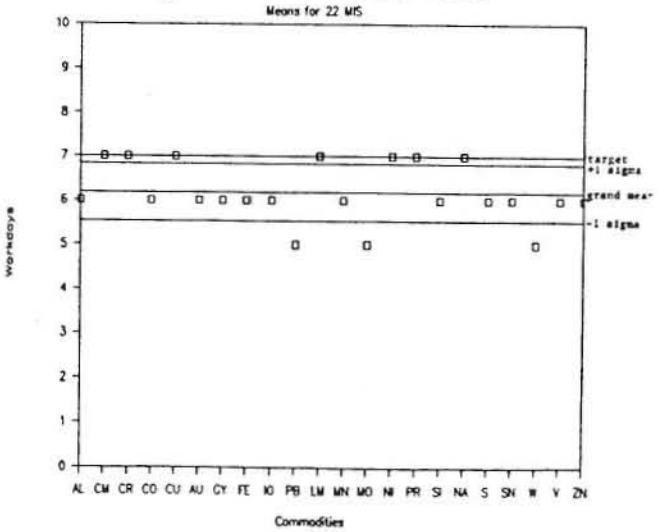


Fig. IV Distributed to the Public



timeliness target, require no action and are not listed in table 3.

ANALYSIS

After all MIS were ranked, by activity phase, for priority attention, historical data were used to construct SPC charts. A control chart is an objective management tool for maintaining control over the behavior of a process, which operates continuously under a relatively stable set of conditions, generally referred to as a "constant-cause system."

Normal variations are inherent in every constant-cause system and they are effected by many relatively minor and unidentifiable factors. They can influence the system somewhat but will not have major impact. However, "unusual," "abnormal," or "identifiable" variations are those caused by one or more major factors not part of the continuing constant-cause system. These factors are called "assignable causes" and have an outstanding impact on the system. It is worthy of management's efforts to detect and eliminate them.

The following steps describe how to construct an SPC chart:

1. Obtain measurable data reflecting the variation of workdays of completion under the constant-cause system for each activity phase of the MIS under study.

2. Establish a standard from the data for each activity phase, i.e., the mean workdays of completion.

3. Accept the normal or usual variation of workdays of completion around their mean value. This is referred as tolerance.

4. Accept a level of "risk"; therefore, the probability of discovering nothing of importance is known in advance.

5. Based upon the level of "risk," establish the upper control limit (UCL) and lower control limit (LCL).

6. Analyze variation from the mean on a continuing basis in relation to these tolerances.

The primary purpose of the control chart is to provide a guide for action to improve the process.

RESULTS

SPC charts were developed for all the surveys identified and ranked for priority attention in managing the MIS publication process. The level of risk was set at 10 per cent in each case. This means that the risk of an assignable cause of variation actually being due to chance is ten percent. For the purposes of this paper, three MIS were selected to serve as examples. They come from the highest priority ranking in each activity phase. Their SPC charts were updated with the most recent timeliness data available, as shown in figures 5, 6, and 7.

IRON AND STEEL SCRAP was selected to represent the **Document Closeout** phase because the timeliness data for the period November 1991 through November 1992 indicated significant improvement in the process. Furthermore, the trend line for this period became stable and the timeliness target was met in 11 of the 12 months.

IRON ORE represents the **Tables Completed** phase because of the improvement in timeliness which occurred from November 1991 through December 1992, particularly during the last half of this period. Because only 2 of 13 months met the timeliness target, however, efforts to improve the process are continuing.

ZINC represents the **Forwarded for Printing** phase because efforts to improve timeliness have been successful. Not only has the trend line become stable but also the timeliness target has been met consistently since January 1992.

CONCLUSIONS

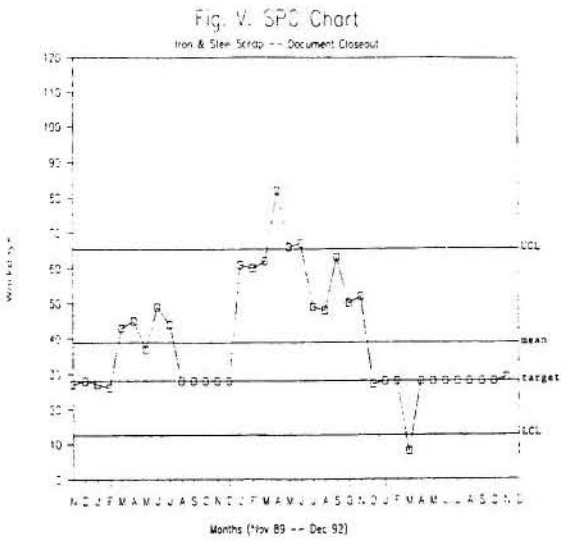
SPC charts assist process managers to focus attention on significant variations from normal conditions and avoid undue concern with the numerous insignificant

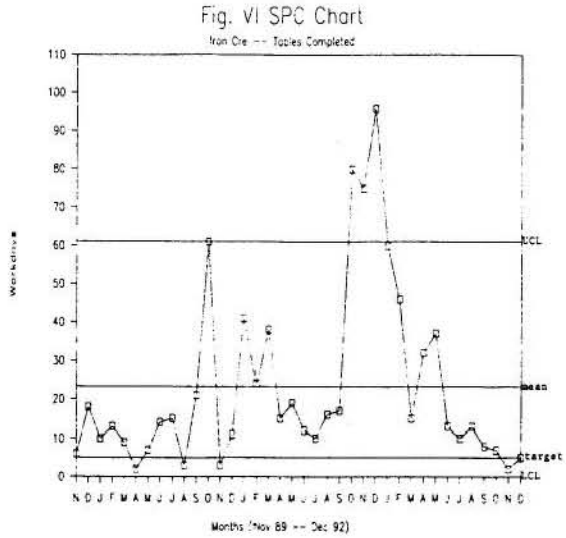
or unidentifiable variations which occur in any constant-cause system. When investigation detects specific causes (assignable cause), appropriate action can be taken to prevent the recurrence of problems.

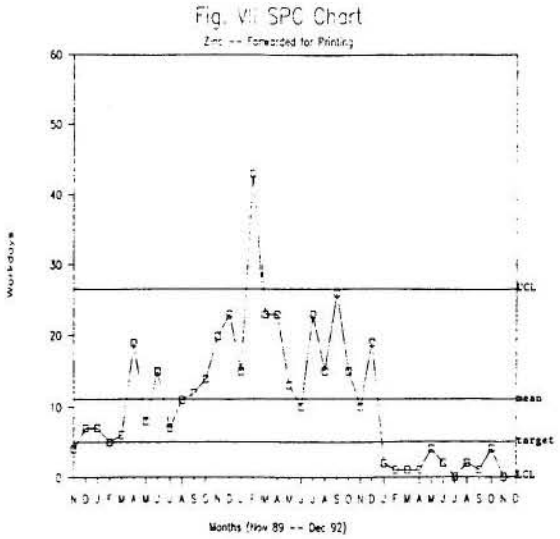
SPC charts are being updated as needed to monitor the MIS publication process. Subsequent monitoring has indicated substantial improvement in the timeliness of many surveys. Updated SPC charts show a decrease in process variation and trend lines heading toward timeliness targets. Several management actions have been effective in improving MIS timeliness. For example,

- a Quality Improvement Team's recommendations regarding interface among organizational units involved in the MIS process resulted in an effective change in the procedures for survey data review,
- a work unit was tasked to implement procedures to assure uniform data processing and table preparation for MIS publications,
- statistical standards were developed, published, and distributed to all employees within the past year,
- statistical positions were upgraded within the statistical organization,
- capable contractor support in processing mineral surveys was instituted.

Since the quality improvement in MIS is an on-going process, the Branch of Statistics and Methods Development will continue to monitor 22 MIS. An annual review of each activity phase will be conducted accordingly when new data become available.







AN EVALUATION OF TAXPAYER-ASSIGNED PRINCIPAL BUSINESS ACTIVITY (PBA) CODES ON THE 1987 INTERNAL REVENUE SERVICE (IRS) FORM 1040, SCHEDULE C

Carl A. Konschnik, Jock R. Black, Richard A. Moore, and Philip M. Steel
Bureau of the Census
Carl A. Konschnik, Business Division, Bureau of the Census, Washington, DC 20233

KEY WORDS: Nonemployer classification, IRS Form 1040 Schedule C (1040-C), Principal Business Activity (PBA) codes, Standard Industrial Classification (SIC) codes, self-classification, Standard Statistical Establishment List (SSEL)

I. Introduction

This paper describes the results of a study designed chiefly to determine the quality of Principal Business Activity (PBA) codes assigned by taxpayers on their 1987 Internal Revenue Service (IRS) Form 1040, Schedule C tax returns. IRS requires that this schedule be filed annually by sole proprietorship businesses to report basic identifying information and characteristics as well as various components of income, expenses, and cost of goods sold. Other aspects of assigning kind of business codes in the economic censuses are also discussed.

For the 1987 Schedule C, taxpayers were asked to select their most appropriate four-digit PBA code from a list of 172 codes and descriptions on the back of the form. They were also asked to write in a description of their principal business or profession, including products made or services performed. Computer files containing the records of selected data from all Form 1040, Schedule C returns for the 1987 tax year were provided to the Census Bureau by the IRS. These records were used to update receipts for employers on the Census Bureau's Standard Statistical Establishment List (SSEL) for all kinds of business and to identify and tabulate nonemployers for the 1987 censuses of retail and services.

A sample of 25,000 Schedule C records which were not matched to an associated employer record on the SSEL and therefore not used to update receipts on the SSEL (these are nonemployers for the most part, although some employers were also included) was selected and mailed a survey questionnaire, Form CB-9924, designed to determine sufficient information to permit accurate kind-of-business coding. The returned survey forms were assigned a Standard Industrial Classification (SIC) code clerically and tabulations were produced which compared the taxpayer-assigned codes (converted to an SIC) with the survey-assigned codes at various coding levels. The impact of coding discrepancies on the nonemployer tabulations in the 1987 retail and services censuses are given in the paper along with estimates of potential coding errors in other economic areas. Other industry coding operations for the censuses were also evaluated, namely, the Census Bureau's automated code assignment based on the written description of the business activity on the Schedule C, and codes assigned by Census Bureau clerks when the taxpayer fails to select a PBA code and the automated system cannot assign one. The paper also contrasts this study with an earlier evaluation of nonemployer kind-of-business coding in the 1977 censuses.

The Census Bureau collects and publishes census estimates in each economic trade area for years ending in 2 and 7. All large and medium-sized firms (based on payroll cutoffs which vary by SIC) and all multi-unit (having more than one place of business or establishment) firms are mailed census forms asking them to supply geographic, kind-of-business, employment, payroll, and receipts data for each establishment in their operation. In addition, in some trade areas, a sample of small employer firms is selected and mailed a census form. Information for an employer firm not mailed a census form and for all census nonrespondents is obtained from administrative data furnished to the Census Bureau by the IRS. Nonemployer statistics are published in the retail trade, service industries, and the construction industries trade areas. Just as for most small employer establishments, the Census Bureau obtains information, such as name, address, Social Security Number (SSN), sales, wages, kind-of-business classification, and other data for nonemployer establishments from the IRS administrative data. These data are gathered chiefly from IRS Form 1040, Schedule C, Form 1065 (partnership return), 1120 (corporation return), and 1120S (S-corporation return).

After obtaining all the necessary information, the Census Bureau tabulates and publishes economic statistics for payroll, employment, and receipts. These statistics are published by kind-of-business for the U.S., states and counties. Although the geographic, employment, payroll, and receipts data are

considered to be reliable, the Census Bureau is concerned that the kind-of-business codes may be incorrectly assigned by some employer and nonemployer tax filers. Establishments mailed a census form are classified on the basis of their self-designation and answers to questions on sales and services. Small employers not mailed a census form are assigned a kind-of-business classification based on one of the Bureau's current surveys, a prior economic census, or other federal agency sources. If no such classification exists, classification is obtained from IRS administrative records. Specifically, all nonemployer kind-of-business codes are based on IRS administrative data.

Although accurate kind-of-business classifications are essential for reliable statistics, the quality of the codes based on IRS administrative records has long been a problem. Indeed, the quality of kind-of-business classification for nonemployer sole-proprietorships from the IRS, Form 1040, Schedule C returns for 1982 was so poor that statistics for nonemployers could not be published in the 1982 censuses. The coding method used for 1982 consisted of a clerk at an IRS Service Center assigning a code based on the taxpayer's written description. In order to reduce clerical work and improve the codes for the 1987 Economic Censuses, the IRS instituted a self-classification system for sole-proprietorship returns (Form 1040, Schedule C). For the 1987 census, the system assigned a kind-of-business code in one of three ways. First, the taxpayer was asked to find a description of his business activity in PART IV of the IRS Form 1040-C, "Principal Business or Professional Activity Codes", and to transcribe the corresponding PBA code to Item B. The taxpayer was also asked to give a written description of his "principal business or profession, including product or service" in Item A. For census processing, if Item B contained a valid PBA code, this code was translated to the corresponding 1987 Standard Industrial Classification (SIC) code. The 1040-C's without valid PBA codes were then subjected to an automated procedure, which assigned the appropriate SIC code to all records with an explicit 20-character description. Finally, clerks reviewed all records which had not previously been assigned a kind-of-business code and assigned the most appropriate code based on the 20-character description from item A. All records which could not be assigned a valid kind-of-business classification were not tabulated in the 1987 economic censuses. This was about 4 percent of the 1040-Cs received from the IRS. This paper examines the quality of the kind-of-business classifications assigned in each phase of the system. While the study was primarily concerned with the quality of the coding assigned to nonemployer administrative records in the retail trade and services industries areas, and to assess its impact on the 1992 censuses, it also examined the quality of the coding assigned to 1040-C records in all trade areas.

II. Results

The following six observations summarize our major findings.

1. For cases which were not assigned a PBA code by the taxpayer, the code assigned by the automated system agreed with the evaluation survey code at virtually the same rate as the agreement between the taxpayer and the evaluation survey code. One can conclude from this that the practice of having the taxpayer provide both a code and a written description of his activities, but keying only the written description if a valid code is not keyed, is relatively effective as well as cost efficient.

2. Across all kinds of business, the rates of agreement between the taxpayer and evaluation survey codes to the nearest percent (in terms of the dollar volume of sales or receipts) are: 89 percent at the industry level; 77 percent at the two-digit SIC level; 68 percent at the three-digit SIC level; and, 67 percent at the four-digit SIC level.

3. For probable nonemployers only, the rates of agreement between the taxpayer and evaluation survey codes to the nearest percent (in terms of the dollar volume of sales or receipts) are: for retail--86 percent at the industry level, 75 percent at the two-digit SIC level, 69 percent at the three-digit SIC level and, 67 percent at the four-digit SIC level; for services, comparable percentages are 94, 78, 72 and 72 percents, respectively.

4. Miscoding of kind of business for nonemployers in the 1987 retail census had little impact on the dollar volume estimates at the total retail level (.01 percent). At the two-digit level, SIC 57 would have been increased by 2.41 percent and SIC 59 decreased by 1.60 percent. For services, the total for all inscopes services in SICs 7 through 8, would have been lowered by .53 percent. Some two-digit SICs were affected more dramatically: SIC 72 would be decreased by 6.0 percent, SIC 76 increased by 6.78 percent and SIC 83 decreased by 12.14 percent.

5. The methods of classification used for the first time for the 1987 census are generally superior to those used before them. It appears that taxpayers make a sincere attempt to code themselves correctly. The codes which they assigned had similar quality to those assigned by the automated procedure and the clerks. In order to improve the quality of the kind-of-business (KB) coding, one must improve the structure of the PBA codes on the 1040-C and educate the taxpayers on how to correctly assign their appropriate code.

6. Comparing codes assigned to the same units by different coding methods inevitably yields some level of disagreement. For example, a comparison of SSEL SIC codes with Bureau of Labor Statistics assigned SIC codes for a sample of Employer Identification Numbers (EIN's) from the SSEL, yields agreements which are only slightly better than the agreement between taxpayer and evaluation survey coding in this study. Moreover, a review of SIC coding comparisons by Statistics Canada further indicates that other systems of assigning kind-of-business codes have similar rates of agreement.

For retail, the improvement from using the self-classification system appears to be across all kinds of business. This is also true for service industries except for a few kinds of businesses. Most of the discrepancy between their coding and the evaluation study's coding occurs because general business activity descriptions are listed to ensure that all activities in a given trade area are covered. As a result, the taxpayer has to make some assumptions in order to assign himself a code. Occasionally, he makes a wrong decision or codes himself incorrectly into an overly general code which is dubbed a "basket" code. The self-classification system does have a major advantage over the older system in that the coder, even though he has limited knowledge of the classification system, has 100 percent knowledge of his business activity. Often enough detail cannot be conveyed in a written description for a clerk, even one well-versed on the SIC codes, to assign an accurate code. The new classification system alleviates this problem somewhat. An in-depth study of 139 taxpayer coded cases where the 1040-C coding disagreed with the CB-9924 coding provided many good examples of the inadequacies of associating a short description of the business with the correct SIC code. In essence, the complexities of the SIC system itself contributes to the problem of assigning accurate codes.

III. The Sampling Frame and Sample Design for the Evaluation Study

The first objective of this study was to evaluate the quality of the kind-of-business classification assigned to all retail trade and services industries records on the file which was referred to as the 1987 Social Security Number (SSN) Universe. This is the universe of all sole-proprietors whose 1040-C does not contain an EIN that matches to an EIN on the 1987 SSEL. Records in the SSN Universe not only contain the sole source of the kind-of-business coding for the tabulation of nonemployers in the quinquennial censuses but also provide a small portion of the coding for the employer tabulations. In cases where a 1040-C record matches to an SSEL record with no SIC code, the SSEL SIC is derived using the 1040-C PBA. Since establishments in the retail and service trade areas may be miscoded into other trade areas, the coding of all SSN Universe records was studied.

For the evaluation study, the SSN Universe was modified so that it was similar to the universe used to tabulate the 1987 censuses. From the original universe of 11,799,149 1040-C records, 1,289,414 records were dropped due to invalid geographic codes or low annualized receipts (since cases with sales or receipts less than $1000 are not included in the census). The remaining records, were then consolidated by SSN, i.e., multiple 1040-Cs with the same SSN were combined into one record. This record was then assigned a major PBA code (the PBA of the 1040-C with largest receipts for all 1040-C's filed under the SSN). From the resulting frame of 9,935,773 combined 1040-C records, a sample of 25,000 was selected and mailed a Form CB-9924 for the evaluation survey.

The sampling frame was stratified by PBA and receipts size, a certainty (wt.= 1) strata determined, and a systematic sample was selected within each noncertainty strata. The sample was designed to meet coefficient of variation constraints of from 1 to 3 percent on the estimate of receipts by kind of business for the evaluation survey derived kind of business. Variances were computed using the random group method.

Of the CB-9924 forms mailed, 17,214 forms were returned and assigned an SIC code. The following section provides additional details of the results obtained by comparing the codes assigned in the survey to those assigned on the 1040-C.

IV. Sources and Quality of the 1040-C Kind-of-Business Coding

In order to quantify the sources of nonemployer coding, the SSN Universe, restricted to establishments in business at the end of 1987, was examined. The evaluation study found that 77 percent of these 8.7 million establishments were assigned a PBA code by the taxpayer. Based on the 20-character description (also provided by the taxpayer), an additional 14 percent were able to be coded using the automated procedure, and 5 percent were coded by clerical review. About 4 percent of the 1040-C's were not able to be assigned a kind-of-business code, and were not tabulated in the 1987 censuses. All percentages are given to the nearest percent.

Table 1 which follows shows the distribution of the sources of the SIC coding for establishments, and, separately, for sales or receipts based on the evaluation survey estimates.

Table 1. Distribution of the Sources of the Kind-of-Business Coding for the SSN Universe

ESTABS IN BUS (END OF 1987)	EVALUATION ESTIMATE OF UNIV SIZE (MIL)	% CODED BY TAXPAYER	% CODED BY AUTOMATED PROCEDURE	% CODED BY CLERKS	% NOT ABLE TO BE CODED
ALL IND	8.7	77	14	5	4
RETAIL	1.2	79	11	7	3
SERVICES	4.0	77	15	4	4
1987 SALES AS RECEIPTS	EVALUATION ESTIMATE TOTAL (BIL $)	% CODED BY TAXPAYER	% CODED BY AUTOMATED PROCEDURE	% CODED BY CLERKS	% NOT ABLE TO BE CODED
ALL IND	356.7	74	15	6	5
RETAIL	76.6	77	10	9	4
SERVICES	107.2	74	18	4	4

Table 2. Percentage Agreement in Terms of Weighted Establishments and Weighted Sales Between the CB-9924 and the 1040-C Coding for All Kinds of Business in the SSN Universe

	UNADJUSTED				ADJUSTED			
WTD ESTABS	ALL	TP	AUTO	CLER	ALL	TP	AUTO	CLER
IND	84	84	83	71	86	87	86	76
2-DIGIT	67	67	67	58	70	71	71	63
3-DIGIT	60	60	60	50	65	65	65	57
4-DIGIT	59	59	60	48	64	65	65	55
	UNADJUSTED				ADJUSTED			
WTD SALES	ALL	TP	AUTO	CLER	ALL	TP	AUTO	CLER
IND	84	85	86	78	89	89	90	84
2-DIGIT	69	69	70	65	77	77	78	74
3-DIGIT	60	60	63	55	68	68	70	64
4-DIGIT	59	59	62	51	67	67	70	61

The CB-9924 and 1040-C kind-of-business codes may disagree because of three reasons. First, the survey evaluation form, CB-9924, may have been incorrectly coded. We will refer to this as "coder error" (in this case we will assume the taxpayer or census coding is correct). Second, the taxpayer filed multiple tax returns (1040-C's, 1065's, and 1120's) and the major source of his gross income described on the CB-9924 corresponded to his 1065 or 1120 or another 1040-C activity. We will refer to this as "response error" (in this case also we assume the taxpayer or census coding is correct). Finally, when no error is found in the CB-9924 coding and the disagreement of the coding cannot be explained in any way, we assume that the census or taxpayer coding is incorrect. This evaluation study attempted to identify and measure this last type of error including measuring its effect on 1987 census estimates.

A systematic sample of 554 CB-9924 forms was examined to determine the magnitude of coder and response error. Coder error was found in 12 (2.2 percent) of these, and response error in another 9 (1.6 percent). From this sample, we roughly estimated the percentage of the disagreement due to coder or response error and computed an adjustment factor to account for these two types of error. Because the number of CB-9924's sampled was small, the same factor was used to adjust estimates based on taxpayer (TP), automated (AUTO), and clerically (CLER) assigned codes. This roughly assumes that coder and response error is the same across trade areas and at all SIC levels. Table 2 shows a comparison of the CB-9924 coding with each method of assigning a code to the 1040-C's in the SSN Universe. Both unadjusted and adjusted figures are shown.

On an adjusted basis, Table 2 shows that of the 96 percent of all 1040-C's from the SSN Universe which were assigned a code, about 86 percent agreed with the CB-9924 coding at the industry level, 70 percent at the 2-digit SIC (or major group) level, 65 percent at the 3-digit level, and 64 percent at 4-digit SIC level. The automated procedure assigned codes which agreed with the CB-9924 codes at about the same rate as the taxpayer, while the clerically assigned coding showed the poorest agreement with the survey evaluation coding. This should not be interpreted as poor performance by the clerical staff, but rather that they were given the most difficult cases to code (namely, the 1040-C's to which the taxpayer neither assigned a valid PBA code nor gave a description of his business activity that was explicit enough for the automated system to assign a classification).

Similarly, on an adjusted basis, for weighted sales, Table 2 shows that of the 95 percent of the dollar volume of sales of cases assigned a code, 89 agreed with the CB-9924 coding at the industry level, 77 percent at the 2-digit SIC (or major group) level, 68 percent at the 3-digit level and 67 percent at the four-digit level. These are virtually the same percentages as for the taxpayer assignments since that is the largest component.

Although separate tabulations were not made for retail and service cases in the entire SSN universe, they were made for the universe restricted only to probable nonemployer cases. Restricting the SSN Universe to probable nonemployers generally yields percentages which differ only slightly from those presented in Tables 1 and 2. Some of these are cited in the results section (II.3.). However, the adjusted numbers for the probable nonemployers (for all kinds of business) in some cells are as much as 3 to 4 percentage points lower than in Table 2. This illustrates the variability in the adjustment factors. Further details are provided in a full version of the report of this investigation which gives extensive tables showing the comparisons of the coding methods. Looking only at retail, the adjusted percentages are only slightly less than the comparable ones in Table 2; for services, the adjusted percentages are almost always better than those in Table 2.

The full version of this report also shows comparisons of evaluation survey and taxpayer codes across SIC trade areas: agriculture, mining, construction, manufacturing, transportation, wholesale, retail, finance, and services. In summary, it appears that the taxpayers have a tendency to code many activities erroneously to services. Also, while there is a good deal of misclassification between retail and wholesale, more (in terms of dollar volume) gets erroneously classified by the taxpayer to retail when the proper classification is wholesale than vice-versa.

V. Effect of the Nonemployer Kind-of-Business Classification on the 1987 Censuses of Retail Trade and Service Industries

The study investigated the effect of nonemployer coding error on the summary statistics for the 1987 Censuses of Retail Trade and Service Industries. For retail trade, it was anticipated that the quality of the nonemployer coding would have had little effect on the published values, since nonemployer sales constituted only 47 billion dollars (or 3 percent) of the retail total of 1.5 trillion dollars. This study confirmed that overall retail sales were underreported by only 0.2 billion dollars (or 0.01 percent). Total sales of all the retail major groups also showed small changes from the published values, with the largest increase of 1.8 billion dollars (2.41 percent) in the Home Furnishings group (SIC 57) and the largest decrease of 2.4 billion dollars (1.6 percent) in the Miscellaneous Retail group (SIC 59).

For services in 1987, nonemployer receipts constituted 96 billion dollars (or 11 percent) of the 868 billion dollar total. This study estimated that service receipts were overreported by 4.6 billion dollars (0.53 percent). Nonemployer receipts comprised at least 15 percent of the total receipts of 4 SIC major groups: (1) 27 percent for Personal Services, SIC 72, (2) 20 percent for Educational Services, SIC 82, (3) 23 percent for the Social Services, SIC 83, and (4) 15 percent for the Miscellaneous Repair Services, SIC 76. For this reason, the census published service receipts showed greater discrepancies at the major group level than were found in retail. Miscellaneous Repair Service receipts (SIC 76) were the most underreported at 1.7 billion dollars (6.78 percent of the 76 major groups total), while Business Service (SIC 73) receipts led the overreporters at 4.9 billion dollars (or 2.59 percent). Personal Services also showed a substantial overreporting of 2.6 billion dollars (6.01 percent) in receipts.

Table 3 shows the net effects on the 1987 census published sales or receipts totals had the evaluation classifications been used.

Table 3. Effect of Miscoding on 1987 Census Sales and Receipts (in Billions of Dollars)

SIC	1987 CENSUS SALES		EVALUATION SURVEY ADJUSTMENTS		
	EMPLOYER SALES	NON-EMPLOYER SALES	CHANGE TO CENSUS	NEW CENSUS TOTAL	PERCENT CHANGE TO CENSUS
RETAIL	1493	47	0.2	1540	0.01
52	81	2	0.7	84	0.85
53	181	1	-0.5	181	- 0.27
54	302	8	-1.0	308	- 0.31
55	333	9	-0.9	342	- 0.25
554	102	3	0.9	106	0.86
56	77	2	0.8	80	0.99
57	75	3	1.9	80	2.41
58	149	5	0.6	154	0.41
59	139	14	-2.4	150	- 1.60
591	54	0	0.0	54	0.01
SERVICE	772	96	-4.6	864	- 0.53
70	52	2	-0.2	53	- 0.32
72	31	12	-2.6	41	- 6.01
73	166	23	-4.9	184	- 2.59
75	51	7	0.0	58	0.17
76	21	4	1.7	26	6.78
78,79,84	58	7	1.2	66	1.89
80	182	14	1.4	198	0.74
81	67	5	0.5	73	0.74
823,4,9	4	1	0.2	6	3.60
83	7	2	-1.2	8	-12.14
87	127	14	-1.3	140	- 0.89
89	4	6	0.3	10	2.84

VI. Comparison of the Quality of Codes Assigned Using the Modified Self-Classification System (1987 Economic Censuses) Versus That of the Codes Assigned Using the Clerical System (Economic Censuses Prior to 1987)

For economic censuses prior to 1987, the kind-of-business codes were clerically assigned to nonemployer records using the written description provided by the taxpayer on the 1040-C. In

1982, Shimberg and Trager [3] reported their evaluation of the coding used to tabulate the 1977 Nonemployer Censuses of Retail Trade and Service Industries. The 1987 evaluation study showed that the self-classification system improves the classification for retail and is quicker and less costly to implement. Table 4 shows the comparison. Because the 1977 study did not consider sales or receipts volume, no comparisons were done for these variables.

Table 4. Comparison of the Agreement of the Evaluation Study Codes With the Administrative Codes Used for the 1977 and 1987 Census for Sole-Proprietor Nonemployer Establishments

ESTIMATE OF ESTABLISHMENT PERCENTAGE AGREEMENTS WITH IDENTICAL CODES AT THE INDUSTRY AND SIC MAJOR GROUP LEVEL

	77 EVAL STUDY		87 EVAL STUDY	
ADMI N SIC	INDUSTRY LEVEL	2-DIGIT LEVEL	INDUSTRY LEVEL	2-DIGIT LEVEL
52	42	36	55	45
53	77	24	92	21
54	87	71	90	68
55	68	64	74	64
56	72	66	75	56
57	49	41	70	56
58	93	84	96	90
59	62	48	79	70
RETAIL	69	56	80	68
70	56	55	75	57
72	87	79	82	63
73	63	50	77	56
75	85	79	86	81
76	85	75	83	77
78	83	58	86	61
79	80	75	90	65
8X	92	84		
80			94	81
81			95	82
82			95	72
83			96	61
86			96	92
87			84	62
89			79	39
SERVICE	79	70	85	66

VII. Evaluation Study Agreement with the 1040-C Coding Versus the SSEL Coding Agreement with the Bureau of Labor Statistics Employer List Coding

The Census Bureau in conjunction with the Bureau of Labor Statistics (BLS) compared industrial classifications between the Bureau's SSEL and BLS' Business Establishment List (BEL). The Bureau sent BLS three files of SSEL records, which BLS then matched to the BEL. For establishments contained on both lists, the two SIC codes were compared and the results reported in Monk, et al. [2].

Of particular interest to the evaluation study are approximately 50,000 single unit SSEL records sent to the BLS from the 1987 Economic Censuses. Approximately 31,000 of these records matched to the BEL. The rates at which the SIC coding agreed was then calculated at the industry, 2-, 3-, 4-digit levels. Table 5 shows that these rates were very similar to the adjusted estimates found for the nonemployer portion of the SSN Universe. See the above cited report for more detailed results of the BEL versus SSEL study.

Table 5. Percentage Agreement of Establishment Coding

LEVEL	BEL VS. SSEL	1040-C CODE VS. EVALUATION SURVEY CODE	
		ADJUSTED	UNADJUSTED
INDUSTRY	85	86	83
2-DIGIT	73	70	67
3-DIGIT	65	64	60
4-DIGIT	60	64	59

A survey of the literature of other studies revealed that our findings were not that unusual. Assigning business activity classification codes is not an exact science and a certain amount of error often occurs. In order to contrast the comparability of our results with the results of two studies performed by Statistics Canada one can refer to the paper by Colledge, Estevao, and Foy [1]. Generally, those studies show coding agreement similar to what we obtained in this study.

VIII. References

[1] Colledge, M.,, Estevao, V., and Foy, P., (1987),"Experiences in the Coding of Administrative Data", Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 529-534.

[2] Monk, H., Raglin, D., Hanczaryk, P., Chapman, D. and Holgado, A., "Report on the Bureau of the Census and Bureau of Labor Statistics Industry Classification Matching Activities", Bureau of the Census, Washington, D.C. (December 1991), Unpublished Memorandum.

[3] Shimberg, M. and Trager, M., "Evaluation of the Use of Administrative Data for Nonemployers in the Retail and Service Censuses", Bureau of the Census, Washington, D.C., (January 1982), Unpublished Memorandum.