USE OF NEW DATA COLLECTION METHODS IN ESTABLISHMENT SURVEYS

William L. Nicholls II, (Retired) and Thomas L. Mesenbourg Jr., U.S. Census Bureau, Stephen H. Andrews, U.S. Bureau of Economic Analysis, and Edith de Leeuw, MethodikA Amsterdam[1]
S.H. Andrews, BEA, 1441 L St. NW, BE-40, Washington, DC, 20230, Stephen.Andrews@bea.doc.gov

ABSTRACT

This paper describes data collection methods of establishment surveys from a 1999 canvass of 615 surveys at 13 prominent government statistical organizations in the U.S. and other countries. Mail remains the most prevalent collection method, but compared with reports from a similar canvass in 1993, mail is are now more likely to be supplemented with other methods, including FAX returns, administrative records, respondent tapes, or electronic methods. The 1990s also have witnessed the growth of newly emerging collection and capture methods for establishment surveys, including optical character recognition, electronic data interchange, and computerized self administered questionnaires. This paper also examines differences in collection methods by survey sample size, frequency of administration, survey content, and organization.

Keywords: **Questionnaires, Mail, Telephone, Electronic, CASIC, OCR**, **EDI**

1. BACKGROUND

In preparation for the 1993 ICES Conference, Anders Christianson and Robert Tortora (1995) completed an international survey of government establishment surveys to determine, in part, the data collection methods these surveys employed. They found that almost half these establishment surveys (48%) used mail-out, mail-back forms as their **only** method of data collection. An additional 32 percent used mail in combination with other methods, giving a total of 80 percent using mail. Other common collection methods included telephone and personal visit interviewing, administrative records, and data files provided by respondents. Early uses of computer assisted and electronic methods also were reported. Organizations in the 1993 canvass were asked additional questions about surveys they chose as their most important. Of these 104 surveys, less than one in 12 had changed their data collection methods in the recent past, but one in six planned to change their collection methods in the next two years, principally by adopting telephone-based, computer-assisted, or electronic methods. The stage appeared set for a decade of change in establishment survey collection methods.

During the 1990s, household surveys (also called social surveys) underwent a massive change in data collection methods. Computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI) almost wholly replaced paper-and-pencil (P&P) household interview surveys at such national statistical organizations as the U.S. Census Bureau, Statistics Canada, Statistics Netherlands, Statistics Sweden and the U.K. Office for National Statistics. (*Cf.* Couper and Nicholls 1998; Clark, Martin, and Bates 1998.) Establishment surveys were generally perceived as changing their methods much more slowly. This paper describes a 1999 canvass of government-conducted establishment surveys undertaken to learn whether P&P mail questionnaires continued their dominance of establishment surveys during the 1990s and how far modern data collection technologies have progressed in replacing the traditional methods used in the 1980s and before**.**

2. THE 1993 AND 1999 CANVASSES OF ESTABLISHMENT SURVEYS

2.1 Definitions

The definition of establishment (or business) surveys followed the approach by Cox and Chinnappa (1995) for the first ICES conference. Establishment surveys are those that study the characteristics, attributes, or activities of organizational entities, such as businesses, farms, schools, government agencies, institutions, or nonprofit organizations. Individuals within these organizations are questioned only as spokespersons for their organizations. Establishment surveys may ask questions about establishments themselves or about organizational entities at other levels of aggregation, such as

[1] *This paper reports the results of research and analysis undertaken in part by staff (and former staff) of the U.S. Census Bureau and Bureau of Economic Analysis. It has undergone a more limited review by the Census Bureau and by the Bureau of Economic Analysis than their official publications. Views expressed are solely the authors' and do not necessarily reflect those of the Bureau of the Census or Bureau of Economic Analysis.*

divisions or enterprises. The term "survey" is used broadly to include economic or establishment censuses, one-time sample surveys, and those conducted continuously or periodically. In this canvass, the term "establishment surveys" also includes the collection of data about establishments from administrative records, from respondent supplied data files or printouts, or by electronic data interchange. Descriptions of the various data collection methods currently available for establishment surveys are explained in a later section. For more detailed descriptions of the newer collection methods, see Couper and Nicholls, 1998, and the Glossary of CASIC Acronyms in that same volume.

2.2 The Sample

The 1999 canvass of establishment surveys lacked the financial, staffing, and linguistic resources of the 1993 canvass (Christianson and Tortora 1995). As shown in Table 1, the 1999 canvass was smaller both in number of organizations and surveys. Efforts to more closely match the samples and questionnaires of the two surveys were forestalled when it was learned that the records of the 1993 study had not been retained. The comparability of the two studies is therefore somewhat uncertain. The 1999 canvass chose a purposeful sample of 13 "prominent national statistical agencies," seven in the United States and one each in six other countries. Except where approvals have been received, participating organizations will not be named. The results are described as applying to "prominent statistical agencies" because the sample represents most of the government agencies whose staff organized the 1993 and 2000 ICES. The most serious omissions are organizations in countries which do not include English as an official first or second language. U.S. government agencies whose surveys typically were conducted by private contractors also are omitted. The results apply primarily to industrialized countries with reliable postal mail systems. Elsewhere, establishment surveys are more commonly conducted by personal visit interviews.

2.3 Comparison of Basic Results

As shown in Table 1, a larger percentage of establishment surveys were reported using mail methods in the 1999 canvass (85.2%) than in the 1993 study (79.9%). Mail apparently remains the predominant collection method of establishment surveys. Two striking differences are nevertheless apparent in Table 1. First, mail was less frequently reported as the **only** collection method in 1999 (31.6%) than in 1993 (47.8%). Second, in the 1999 canvass, mail was much more often used in combination with other collection methods. The mean number of collection methods per survey increased from 1.50 in the 1993 study to 1.93 in the 1999 canvass. The major change was the use of mail combined with methods other than telephone or personal interviewing, from 2.0% in 1993 to 30.7% in 1999. The

Table 1 – Comparison of the 1993 and 1999 Canvasses of Establishment Surveys

	1993 Study		1999 Study	
Source of Data				
Base year of reference	1992		1998	
Number of countries included	16		7	
Number of organizations responding	21		13	
Number of surveys reported	1,387		615	
Response rate of organizations	75%		100%	
Collection Methods	**Percent**	**Cumulative Percent**	**Percent**	**Cumulative Percent**
Mail only	47.8%	47.8%	31.6%	31.6%
Mail with telephone or personal visit	30.2	78.0	22.9	54.5
Mail with other methods, e.g., FAX returns	2.0	80.0	30.7	85.2
Telephone or personal visit without mail	10.3	90.3	7.8	93.0
Other methods or combination without mail	9.7	100.0	7.0	100.0
Mean number of methods per survey	1.50		1.93	

methods now increasingly used with mail include facsimile (FAX) returns, respondent supplied tapes or printouts, computerized self-administered questionnaires, touch tone data entry, and electronic data interchange. In some surveys, the telephone or personal interviews were used in addition to mail and these new methods.

3. USE OF INDIVIDUAL COLLECTION METHODS

Table 2 summarizes the current usage of individual data collection methods (and their associated data capture procedures) for establishment surveys by percent of surveys, percent of annualized sample cases using the method, and number of organizations in the 1999 canvass. It also reports the median year of first use among organizations, and the number of organizations who considered using a method but did not implement it. We will discuss these organizational results below after introducing the individual methods and their usage measured by survey and by sample cases.

Mail-based Methods.----By 1999, mail surveys took a variety of forms. The most common remained the P&P mail out, mail back questionnaire with data capture by key entry. This traditional method was employed by 79.0% of all establishment surveys in the canvass, typically in conjunction with additional methods. A slight variation is the leave (with respondent) and mail back questionnaire used in 1.6% of surveys. Newer approaches to mail surveys employed a P&P mail out, mail back form intended for scanning, with data capture by optical character recognition (OCR) or by keying from an scanned image of the form, used by 3.9% and 1.0% of the surveys, respectively. (Some surveys now scan P&P forms for archiving without necessarily keying from the image.) Finally, a few surveys employed mail out, mail back forms designed for optical or electronic mark sense data capture. Mark sense is found primarily in education surveys where it is used not only for individual student achievement tests (out of scope here) but also to capture administrative data about their schools. Together, these various mail back methods were employed in 85.1% of all establishment surveys in the 1999 canvass.

Survey count, as a measure of usage, has its weaknesses. First, a method may be used in a survey for only a small percentage of its cases. Second, organizations differ widely in the way they count establishment surveys. A survey of manufacturers consisting of 40 sub-surveys, one for each type of manufacturing, may be counted as one survey or 40, adding either 1 or 40 to the number of surveys using its methods. In the 1999 canvass, organizations were encouraged to group surveys wherever possible; and the investigators imposed grouping on the surveys of one organization reported in an atypically disaggregated form. This partially accounts for the smaller number of surveys in the 1999 canvass compared to 1993.

An alternative approach is to measure usage by cases, that is, by the proportion of the annualized sample completed by each method. An annualized sample was calculated for each reported survey based on its sample size and frequency of administration. For example, a monthly survey of 1,000 per month has an annualized sample of 12,000, while a census of 50,000 conducted every five years has an annualized sample of 10,000. A survey's annualized sample size was then multiplied by the proportion of its returned sample reported for each method to allocate that sample among its various methods.

This approach also has its disadvantages. Not all survey organizations keep records, or can easily guess, the proportion of each survey's completed sample attributable to its various methods. Extensive telephone and e-mail follow up was required to prompt for this information or to guide imputation when necessary. Second, results based on percent of cases can be overwhelmed by a few very large surveys.

When the 615 surveys in the 1999 canvass were arranged by annualized sample size, the first 610 had a combined annualized sample of nearly 16 million cases; the remaining five surveys had a total annualized sample approaching 60 million cases. In analyses based on percent of cases, these five surveys, all with annualized samples over 1million are excluded. The collection methods these five surveys principally employed, in percent of cases, were electronic data interchange (62.4%), administrative records and forms (24.0%), P&P mail (7.0%), and touchtone data entry (5.0%).

Excluding these very large surveys, traditional P&P mail out, mail back or leave and mail back methods accounted for 54.0% of the annualized cases, while scanning the mailed forms, with OCR or keying from images, accounted for 10.6%. Because scanning capture methods tend to be used with larger surveys, it appears more significant in counts of cases than in counts of surveys. Since scanning was first typically used in the 1990s, its sizable contribution to mail survey data capture is a relatively new development. In total, mail methods provided 66.4% of all establishment survey cases.

Telephone-based methods.---- These methods were used in 44.1% of the surveys and accounted for 22.0% of the establishment survey cases. They include the P&P telephone interview and computer assisted telephone interview (CATI), both commonly used before 1990. Perhaps the most rapidly growing new collection method for establishment

Table 2 – Usage of Individual Data Collection Methods: 1999 Canvass					
			Organizations		
Collection Methods	Percent of Surveys Use	Percent of Sample Cases[1]	Median Year of First Use	Number Use Method	Number Considered Method
Base of Figures	N = 615	N=15.7m[1]	N = 12	N = 13	N = 12
Mail-based Methods					
P&P mail out, mail back, data keyed form	79.0%	53.9%	*	12	*
P&P leave and mail back form	1.6	0.1	*	1	*
Scanned form with OCR	3.9	7.8	1992	4	6
Scanned or imaged form, keyed from screen	1.0	2.8	1997	1	8
Mark sense form	0.2	1.8	1992	1	4
Any mail method	**85.1%**	**66.4%**	*	12	*
Telephone-based Methods					
P&P telephone interview	21.6%	7.8%	Pre-1990	10	1
Computer assisted telephone interview CATI	12.7	9.5	Pre-1990	7	1
FAX (facsimile) return	19.3	3.2	1990	9	1
Touch tone data entry (TDE)	2.1	1.5	1991	4	4
Any telephone methods	**44.1%**	**22.0%**	Pre-1990	11	
Personal Visit Methods					
P&P personal visit interview, observation,	14.5%	3.8%	Pre-1990	7	2
Computer assisted personal visit, observation	0.5%	0.6	1996	3	4
Any personal visit method	**15.0%**	**4.4%**	Pre-1990	8	
Records or Respondent Tape, Printout					
Administrative records or forms	4.4%	2.3%	Pre-1990	7	4
Respondent prepared tape, disk, printout	10.1	2.1	Pre-1990	7	4
Either of above	**13.3**	**4.4%**	Pre-1990	10	*
Electronic Methods					
Computer self-admin. questionnaires	6.5%	0.3%	1992	5	4
Web or e-mail surveys	2.6	0.2	1996	5	6
Electronic data interchange (EDI)	3.9	2.5	1995	5	4
Any electronic method	**11.5%**	**3.0%**	1995	8	
Reminder methods					
Mail reminder	36.4%	**	*	9	*
Telephone or CATI reminder	34.5	**	Pre-1990	11	1
FAX reminder	5.4	**	1994	3	4
E-mail reminder	0.2	**	1998	1	5
Any reminder method	**60.2%**	**	*	**12**	*

* Not asked or unknown. ** Inapplicable

[1]Column excludes five surveys with annualized sample size of 1,000,000 or more. Base of percentages is 15,677,155.

surveys, especially in the United States, is the facsimile (FAX) return of the survey questionnaire. FAX returns were reported for 19.3% of the surveys, and this is probably an underestimate. Several major data collection organizations, such as Statistics Canada, now routinely include a FAX number with all mailed establishment questionnaires and do not necessarily keep counts of FAX returns separate from the processing stream for mail returns. Although FAX returns are now accepted in many establishment surveys, they are typically an optional method of reply for respondents. They do not account for a large percent of the returns, only 3.2% as shown in Table 2. Only one survey reported using FAX

returns as its principal method of data collection. Touch tone data entry, in which respondents answer a mailed form by dialing the statistical organization and entering their information using the touchtone pad on their telephones, is infrequently used for establishment surveys, although this is one of the methods used by two of the five largest surveys.

Personal visit methods ---- These methods have long been used by establishment surveys in the areas of health, education, agriculture, and housing construction, where personal contacts, personal observation, or recording of records from local government offices are necessary. Personal visits by specially-trained field staff also may be employed as a follow up method for mailed questionnaires. Personal visit methods were employed by 15.0% of the surveys and account for 4.4% of their returns. P&P remains the most common mode of personal visit methods with only infrequent use of laptop or notebook computers for CAPI or its equivalent in observation or transcription of records.

Records, Respondent Tapes.----Administrative records have a long history as a source of data on establishments. Administrative records were employed in 4.4% of the canvass surveys and accounted for 2.3% of the cases in surveys with less than 1million annual sample size. They also were extensively used in the five largest surveys.

When large enterprises with many separate establishments are asked to respond to survey requests, some reply by providing a data tape, disk, or printout of the requested (or similar) data in lieu of the completed forms. Statistical organizations usually discourage such replies because *ad hoc* data files require extra work to interpret and to merge into the survey response data base. However, this method does accommodate respondent preferences and obtain information from large enterprises that might not otherwise reply. They were used in 10.1% of all surveys, but accounted for only 2.1% of their returns. Statistics Netherlands (Ypma, Willeboordse, and Keller 1997) is attempting to encourage and systematize such replies by working individually with large enterprises to establish common standards for the data to be reported while leaving the enterprise free to choose the medium and software employed.

Electronic Methods.----Electronic data collection methods were employed in 11.5% of the canvass surveys and accounted for 3.0% of the cases. One form of electronic data collection is the computerized self administered questionnaire (CSAQ) or computer assisted self interview (CASI). An electronic questionnaire is sent to the respondent, usually on floppy disk, which the respondent installs on his/her computer and then answers. The completed data are then sent back to the statistical organization by disk or modem. CSAQ methods were used in 6.5% of the establishment surveys in the 1999 canvass, but accounted for only 0.3% of the cases. Use of electronic mail or the world wide web in establishment surveys is even newer, accounting for 2.6% by survey and just 0.2% of the cases. However, many of the statistical organizations in the 1999 canvass apparently believe that web or e-mail establishment surveys will become increasingly important in the future; five agencies had already used them and six others had considered their use.

Electronic data interchange (EDI) involves the direct transfer of establishment data from a respondent's computerized data base to that of a statistical organization. By automating the extraction of data from respondent records, EDI increases the timeliness of survey reporting and reduces both its costs and respondent burden. In practice, however, establishments often are reluctant to devote the time and resources necessary to prepare for EDI statistical reporting, and most surveys employing EDI seem to require continued use of more traditional methods for much (or at least part) of their sample. The usage of EDI ranges from surveys that employ it for small numbers of respondents who request it, to very large surveys, especially in foreign trade, that use it as their primary collection method. Some survey organizations, especially Statistics Netherlands (Ypma, Willeboordse, and Keller 1997), regard it as the most promising method for the future.

Reminder Methods.----Table 2 also includes usage figures on nonresponse reminder methods. Reminder methods are used to encourage establishments to respond via other collection methods, not to collect data themselves. Hence, we cannot attribute a portion of the sample cases to them. At least one reminder method was used in 60.2% of the surveys. The familiar mail and telephone reminders were used in 36.4% and 34.5% of the canvass surveys, respectively. The newer options of FAX and e-mail reminders were used less frequently. (In the 1999 canvass, of the 194 surveys represented in Table 1 as using only mail questionnaires, 168 prompted response with telephone, mail, or other reminders. Only 26 surveys (4.2% of the total) relied exclusively on mail forms without reminders.) Although Christianson and Tortora (1995) did not mention the use of nonresponse reminders in their report, surveys they classified as relying exclusively on P&P mail collection may actually have used telephone or other reminders to encourage returns.

Survey Organization Use.----Table 2 also presents information on the number of organizations using each method, the median year of first use among those organizations employing it, and the number of organizations who considered the use of each method but are not now using it. The year of first use was not asked for specific years before 1990 or for such traditional methods as the P&P mail questionnaire or mail reminder. "Considering" was defined as taking steps to learn enough about a method to make an informed judgment about its possible use -- not just casually thinking about it.

Information for the 1999 canvass of establishment surveys was sometimes obtained from a single source high in the organizational structure, but more frequently our forms were distributed to the division, branch, or survey level where the details of survey operations were best known. About 60 organizational subunits responded; and questions about year of first use and methods considered were typically asked at that level and cumulated to an organizational summary. Because one organization was unable to answer these questions, the organizational sample size is 12 for those items. Despite these limitations, the reports of median year of first use suggest that many of the current collection methods of establishment surveys were first employed after 1990. These methods include scanning with OCR or keying from images, FAX returns, touch tone data entry, CAPI, CSAQ, web or e-mail surveys, EDI, and reminders by FAX and e-mail. The reports on methods considered by those not using them, suggest that current attention is especially focused on scanning with or without OCR, web or e-mail surveys, and e-mail reminders.

4. COMBINATIONS OF COLLECTION METHODS

The seventeen basic collection methods and four reminder methods listed in Table 2 may be employed by establishment surveys in many combinations. One method of analyzing these combinations, used in Table 3, is to classify the surveys by their use of the principal forms of mail methods and identify the other methods employed with them. The first column in Table 3 describes surveys which used P&P mail or leave and mail back methods with standard key entry data capture. The second column describes surveys which used mail methods with OCR, imaging, or mark sense data capture. For brevity these first two groups will be labeled by their principal methods – P&P mail and OCR mail. Surveys which used both P&P mail and OCR mail are combined with the latter. The third column describes surveys which did not employ any mail methods.

The P&P mail surveys obtained 73.6% of their sample data by P&P mail methods, while the bulk of the remainder included FAX returns, P&P telephone, and CATI. As shown at the bottom of Table 3, these surveys employed a mean of nearly 2 (1.94) basic collection methods per survey, that is P&P mail and one other. However, 36.0% of the P&P mail surveys relied on mail as their sole basic data collection method, while the remainder employed a mean of 3.01 basic collection methods per survey, that is P&P mail plus two others. These additional methods did not account for very large proportions of the returns.

Table 3 – Collection Methods by Annualized Sample: Combinations With and Without Mail

Collection Methods	Mail Methods		No Mail Used	Total
	P&P Mail or Leave Mail Back	OCR Mail, Imaging or Mark Sense		
Annualized sample size = Base of percentage	11,343,569	2,223,408	2,110,178	15,677,155
P&P mail or leave and mail back	73.6%	4.8%	- %	53.9
OCR mail, imaging, or mark sense	-	86.5	-	12.3
Facsimile return	4.0	1.1	1.0	3.1
P&P telephone	9.6	0.1	6.3	7.0
CATI	7.0	3.2	29.7	5.5
Personal visit or CAPI	0.9	0.0	27.6	0.7
Admin. records or respondent tape, disk, printout	2.0	0.6	21.7	1.5
Electronic data interchange	0.6	1.2	13.7	2.5
Other electronic (CSAQ, web, e-mail) and TDE	2.2	2.4	-	2.0
Total of percentages	100.0%	100.0%	100.0%	100.0%
Mean collection methods per survey	1.94	1.77	1.24	1.82
Mean reminder methods per survey	0.81	1.53	0.21	0.76
Base of means	489	30	91	610

[1]Percentages exclude five surveys with annualized sample sizes of 1,000,000 or more.

The OCR mail surveys obtained 86.5% of their sample data by OCR, imaging, or mark sense methods and used P&P mail, and CATI for most of the remaining portions of their samples. The OCR mail surveys made more frequent use of reminder methods, averaging 1.53 reminder methods per survey, than the P&P mail (0.81) and no mail methods

(0.21). The principal advantages of OCR, reduced data capture costs and quicker processing, are maximized when virtually all sample cases have their data captured by OCR. Additional forms of data collection usually mean additional paths of data capture and more complexity in integrating data flows. From a processing perspective, it is more efficient to use reminder methods to encourage additional OCR returns than to add other forms of data collection and capture.

Establishment surveys that did not use mail methods principally employed manual or CATI telephone interviews, personal visits, administrative records, or EDI. These are mostly surveys where time constraints rule out mail methods, or where personal contact or observation are necessary, or where traditional surveys may be replaced by administrative records or EDI. It is interesting to note that such modern methods as CSAQ, TDE, and web surveys occur only with establishment surveys which also use P&P or OCR mail. They supplement or complement mail methods.

5. VARIATIONS BY SURVEY ORGANIZATION AND SURVEY DESIGN

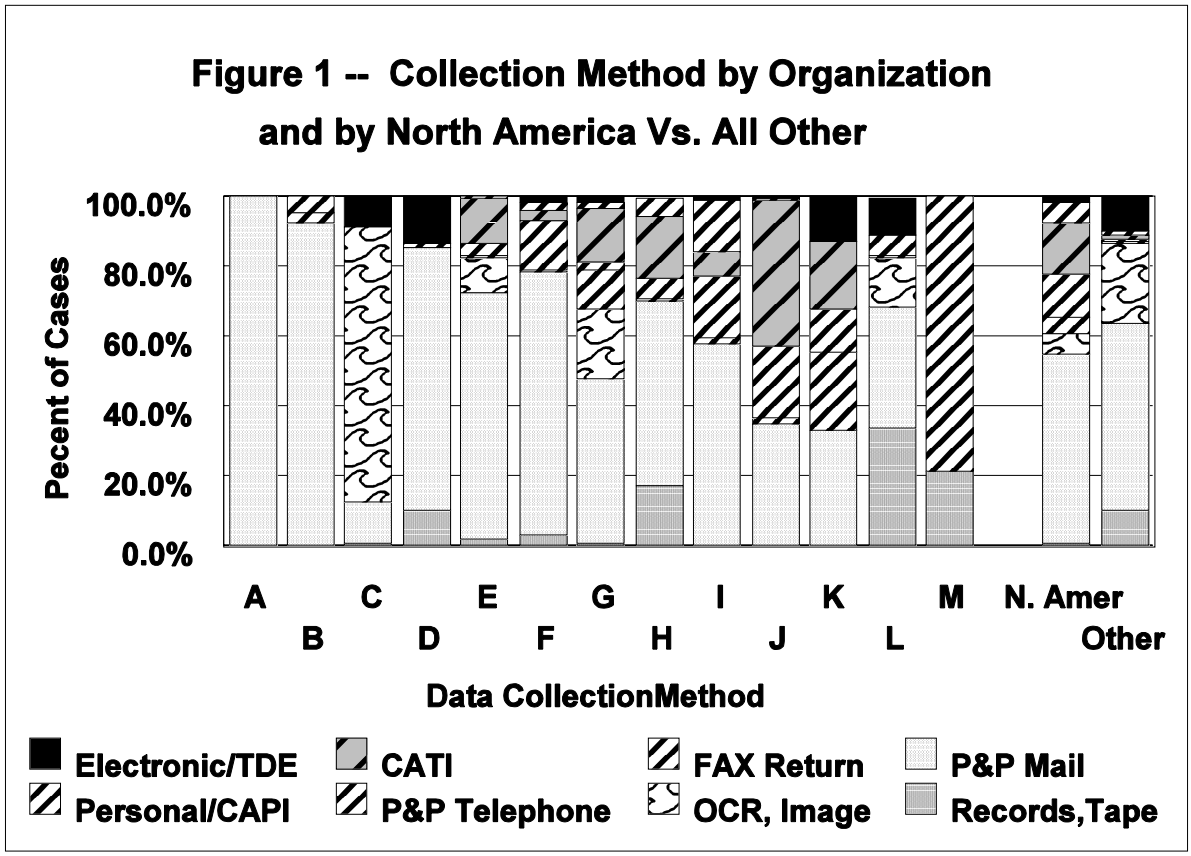
Figure 1 and Tables 4, 5, and 6 examine variations in usage of basic data collection methods by survey organizations, continent, sample size, frequency of survey administration, and survey content. All employ the same groupings of methods, which are most fully described in Table 4. They all measure usage in sample cases and all omit the five largest surveys with annualized sample sizes of 1million or more.

Survey Organization.----As shown in Figure 1, the 13 survey organizations in the 1999 canvass varied widely in the collection methods they employed. Some relied exclusively or almost exclusively on P&P or OCR mail methods. One organization relied exclusively on personal visit methods and administrative records. Although many organizations used P&P telephone or CATI methods, they varied substantially in the portion of cases accounted for by these methods. The organizations also varied greatly in the proportion of their sample data obtained from administrative records and respondent supplied tapes, disks, and printouts. Eight of the 13 organizations obtained measurable amounts of their data by one or more of the electronic reporting methods (EDI, TDE, CSAQ, and web or e-mail) or by TDE.

In part, organizational variations in collection methods may reflect their differing subject matter emphases or the types of surveys they principally conduct. These organizational differences also undoubtedly reflect the organizations' traditions, priorities, and established resources. Investments in printing and mailing equipment, in scanning and OCR technology, in centralized telephone or CATI facilities, in a decentralized field staff equipped with CATI/CAPI laptops, or in the hardware and software necessary for CSAQ, TDE, Web surveys, or EDI represent major institutional commitments. These commitments undoubtedly influence and constrain decisions about the collection methods of the organization's individual surveys. The collection methods used by different divisions within the larger survey organizations are also highly diverse, probably reflecting subject matter differences as well as divisional traditions and commitments.

North America vs. All Other.----The last two columns of Figure 1 contrast the establishment survey collection methods of survey organizations in North America (U.S. and Canada) with those of organizations located on other continents. Both obtained about the same percentage of their cases by P&P mail, but survey organizations in North America obtained proportionally more cases by FAX returns, P&P telephone, CATI, and personal visits. Survey organizations elsewhere obtained larger proportions of their data from administrative records, OCR mail, and electronic or TDE methods. The importance of OCR mail and EDI outside North America may partly reflect laws in Europe and elsewhere limiting the number of hours per day an employee may be assigned key entry tasks. It should be reemphasized that the results in Figure 1 are based on percentages of sample cases in surveys with under 1 million annual sample size. Among these surveys, 46 North American establishment surveys used electronic or TDE methods compared with 31 surveys from other continents, but the latter collected data on four times as many cases.

Sample Size.----Table 4 demonstrates less variation in collection methods by survey size than might be anticipated. Smaller surveys displayed about the same mix of collection methods as larger surveys. Administrative records, P&P mail, FAX returns, P&P telephone, CATI, and electronic or TDE methods are reported for at least some surveys at almost every size category. There are some apparently meaningful variations by size. The usage of OCR mail increases with sample size, while usage of P&P telephone and CATI is most common in surveys of moderately large size, that is with annualized samples between 10,000 and 249,999.



Survey Frequency.----As shown in Table 5, the frequency of survey administration is a much better predictor of survey collection methods. The use of administrative records and respondent prepared tapes or printouts is most prevalent in annual surveys. Mail methods are most common for surveys conducted quarterly, annually, or less often, possibly because time is required to prompt for and receive mail replies. P&P telephone and CATI methods are more common for surveys conducted monthly, weekly, or continuously, where they are often the primary means of hastening delinquent returns, although they also are used in one-time surveys. Electronic or TDE methods also are most likely to make their contributions to monthly or weekly surveys.

Survey Content.----Table 6 examines survey collection methods by survey content, which was assessed from the survey name, survey sampling unit, and organizational subunit responsible for its management. Survey content provides some hints to collection method. Administrative records are most prevalent in surveys of governments, educational institutions, and services. Personal visit methods are common in surveys of health, agriculture, education, and construction. Electronic and TDE methods are very prominent in surveys of international trade. Reasons for other differences shown in Table 6 may be due to survey frequency, survey design, organizational traditions, or respondent burden.

6. SUMMARY AND DISCUSSION

The 1999 canvass of establishment surveys was undertaken in part to measure the extent of an expected transition from mail surveys to more modern electronic methods of data collection for government establishment surveys. Eighty-five percent of the 1999 establishment surveys used mail methods, and mail methods accounted for two-thirds of the establishment survey returns. The major changes observed since the previous 1993 canvass were that a greater number of surveys are using mail in combination with other data collection methods and a change in the way data from mail surveys is captured. A significant and growing minority of large mail surveys have changed from key entry data of paper forms to OCR or to keying from images obtained from scanning.

Why have mail methods continued to represent the major share of establishment survey data collection? In their 1995 chapter, Christianson and Tortora speculated that the popularity of mail methods was a consequence of two factors. The first was the perceived low cost of mail relative to other methods. The second was their adaptability to the

Table 4 – Collection Methods by Annualized Sample Size[1]

Collection Methods	Annual Sample Size					
	Under 500	500-1,499	1,500-9,999	10,000-49,999	50,000-249,999	250,000 or more
Annualized sample size (000s)	33.0	69.6	702.1	2,514.5	7,270.2	5,6067.2
Admin. records or resp tape, etc	9.2%	9.0%	4.5%	2.2%	3.4%	6.9%
P&P mail or lv. & mail back	67.8	64.6	58.6	52.4	45.0	66.5
OCR mail, imaging, mark sense.	0.0	2.1	3.6	6.8	16.0	11.1
Facsimile (FAX) return	3.9	3.7	5.4	3.6	1.1	5.8
P&P telephone	5.8	7.8	6.8	10.0	8.3	6.2
CATI	6.1	6.3	7.9	16.3	12.6	2.1
Personal visit or CAPI	3.2	1.9	9.9	4.2	6.8	0.3
Electronic or TDE[2]	4.0	4.6	3.3	4.5	6.8	1.1
Total of percentages	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 5 – Collection Methods by Frequency of Survey Administation[1]

Collection Methods	Frequency of Survey Administration						
	One Time	Every X Years	Annual	Quart-erly	Monthly	Weekly	Contin-uous
Ann.sample size (000s)	596.7	1,781.1	4,203.1	4,613.1	3,900.2	444.1	138.8
Admin. records, resp. tape	0.0%	0.4%	12.0%	3.5%	0.6%	0.2%	0.0%
P&P mail or lv. & mail	62.6	57.7	53.7	66.6	41.2	25.4	2.0
OCR mail, imaging, etc.	2.2	31.9	14.4	8.0	9.4	0.0	0.0
Facsimile return	0.1	0.1	0.6	5.6	3.0	22.5	6.3
P&P telephone	11.2	0.7	5.9	6.9	10.7	34.4	91.7
CATI	23.9	7.5	7.3	5.1	13.0	9.3	0.0
Personal visit or CAPI	0.0	1.7	5.1	2.4	8.7	0.0	0.0
Electronic or TDE[2]	0.0	0.0	1.0	1.9	13.4	8.2	0.0
Total of percentages	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 6 – Collection Methods by Survey Content

Collection Methods	Survey Content: Selected Categories							
	Intern'l Trade	Govern ments	Estb. in general	Manu-facturg	Agri-culture	Bldg. Constr.	Educa-tion	Service Indust.
Ann.sample size (000s)	631.9	937.5	2.412.9	1054.7	3,832.7	1,091.0	118.3	1,757.2
Admin. records, tape or	0.1%	15.0%	1.3%	1.0%	2.9%	0.9%	23.5%	16.4%
P&P mail or lv. & mail	47.7	82.1	49.9	53.9	54.4	65.0	50.4	46.0
OCR mail, imaging,	7.1	0.5	28.2	11.8	0.0	0.3	0.0	6.8
Facsimile return	0.1	0.1	4.5	8.4	1.9	4.1	0.4	1.8
P&P telephone	0.3	0.2	4.2	3.9	15.4	6.3	4.4	14.2
CATI	0.0	0.0	6.5	15.8	12.0	18.5	13.0	11.5
Personal visit or CAPI	0.0	1.2	1.4	0.1	12.6	4.7	4.3	1.3
Electronic or TDE[2]	44.7	0.9	4.0	5.1	0.8	0.2	4.0	2.0
Total of percentages	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

[1]Percentages excludes five surveys with annualized sample sizes of 1,000,000 or more.
[2]Electronic/TDE methods include EDI, TDE, CSAQ, and web or e-mail.

collection of data from organizational records. Paper forms facilitate the entry of data in any order available from organization records; and paper forms are easily routed around an organization when information is required from different offices. No special equipment or technical skills are required to complete a paper form, and the completed

forms are easily reviewed by managers and copied to maintain a record of the submitted information. With these benefits, it is not surprising that mail P&P forms retain a major role in collecting data from establishments.

Alternative data collection methods have long been available to meet the specialized needs of establishment surveys not readily provided by mail collection methods. Telephone interviews may be used to help respondents through complex questionnaires, prompt delinquent replies, and speed the completion of the field work. Frequently, however, telephone methods are used only for that minority of the sample experiencing difficulties in completing the survey, while mail continues for the majority. Personal visits methods may be used when more personal contact is needed for motivation or to permit observation or transcription of data from local records, but since both telephone and personal visit methods are labor intensive and generally more expensive per case than mail methods, they are used sparingly. The transition to computer assisted interviewing methods (CATI and CAPI) may add new benefits in concurrent data entry, interactive edit checks, and faster turnaround between administrations but does little to change the relative strategic advantages and disadvantages of mail, telephone, and personal visit. Administrative records or EDI can sometimes replace traditional survey data collection, at least for a part of the sample, but the situations where these opportunities exist seem limited.

Some new establishment survey data collection methods may have evolved because respondents encouraged them or find them more convenient than traditional mail methods. They include FAX returns, respondent prepared tapes, disks, or printouts, web or e-mail surveys, touch tone data entry, and CSAQ. However, not all respondents prefer these alternatives and not all respondents have the equipment, knowledge, or interest to employ them. Even when these methods are employed, mail forms are still required for at least part of the sample (and often most of the sample) as a low cost option that virtually all respondents can easily understand and use. Many of the newer methods become supplements to mail data collection procedures rather than replacements for them. This increases the number of methods per survey but does not reduce the general prevalence of mail methods . If there is a harbinger of long run changes in establishment surveys, it probably lies in the very largest surveys and their choices of EDI, administration records, and mixed combinations of several traditional and modern methods for data collection. In the next five to ten years, mail methods will likely remain an important part of establishment survey data collection.

The remaining papers in this session (Clayton *et al.*, Parent and Jamieson, and Keller and Willeboordse, following in these *Proceedings*) describe the historical, current, and planned development of new establishment survey methods at three statistical agencies well known for their innovations in the survey field. They include developments after 1998, the reference year of the 1999 canvass. Without minimizing their significant contributions and their likely influence on the establishment survey methods of the future, it is important to recognize that these developments were occurring at a time when almost all governmental statistical agencies were still relying on P&P mail forms for the overwhelming majority of their data collection from establishments.

REFERENCES

Christianson, A. and R.D. Tortora (1995), "Issuesin Surveying Businesses: An International Survey," in B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, and Philip S. Kott (eds.), *Business Survey Methods*, New York, Wiley, 1995, pp. 237-256.

Clark, C.Z.F., and J. Martin (1998), "Development and Implementation of CASIC in Government Statistical Agencies," in M.P. Couper, R.P. Baker, J.Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls, and J.M. O'Reilly, (eds.), *Computer Assisted Survey Information Collection*. New York: Wiley, 1998, pp. 62-84.

Couper, M.P. and W.L. Nicholls (1998), " The History and Development of Computer Assisted Survey Information Collection Methods," in M.P. Couper, R.P. Baker, J.Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls, and J.M. O'Reilly, (eds.), *Computer Assisted Survey Information Collection*. New York: Wiley, 1998, pp. 1-21.

Cox, B.G. and B.N. Chinnappa (1995), "Unique Features of Business Surveys," in B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, and Philip S. Kott (eds.), *Business Survey Methods*, New York, Wiley, 1995, pp. 1-17.

Ypma, W., A. Willeboordse, and W. Keller (1997), "EDI in the Collection Statistical Data: an Introduction," in J. Pannekoek (ed.) *Netherlands Official Statistics*, Special Issue, *EDI: The State of the Dutch Art*, V. 12, Voorburg: Statistics Netherlands, 1997.

THE USE OF CAI FOR THE COLLECTION OF BUSINESS SURVEYS: IN STATISTICS CANADA

Guy Parent, Statistics Canada
Rolly Jamieson, Statistics Canada
G.Parent, 9C4 Jean Talon, Tunney's Pasture, Ottawa, Ontario K1A 0T6, pareguy@statcan.ca

ABSTRACT

In 1992 we participated in the Annual Research Conference, at which time we outlined our goals and plans for the use of CATI, CAPI and other technologies in the collection of Business Surveys. The paper will review the methods that we have used over the past 7 years including CATI, CAPI, Electronic Reporting, Internet Web data collection and our experiment with touch-tone. It will describe the various CATI development approaches that are required for Annual Surveys Vs Monthly Surveys, as well as the impact of the increased use of on-line-editing that is employed to improve the quality of the data collected. The paper will address the CAI approach to frame updates, the use of on-line coding tools, the pre-contact method, tailored questionnaires and the problems related to multi-contacts establishments. The results of the Workplace and Employee Survey, a survey that has both an establishment questionnaire and a separate employee questionnaire using mixed mode collection (Mail, CAPI, CATI) within one survey will also be covered. The paper will conclude with some of the issues related to the increased and it's possible impacts.

Keywords: CAI, CATI, CAPI, Electronic Data Reporting (EDR)

1. INTRODUCTION

Data collection activities at Statistics Canada have undergone profound changes during the last decade. The emergence of technology in all aspects of data collection, has radically changed the way we plan, design, and execute such activities. Computer Assisted Interviewing (CAI) has now become a standard way of doing business. The primary purpose of this paper is to discuss the main benefits that have accrued and the challenges that emerged during the planning and implementation phases

CAI, as a new approach held many promise most of which were realized and are now being exploited to the fullest. The transition in adapting and deploying these new collection technologies presented a challenge and opportunity for innovation which surpassed other data collection initiatives at Statistics Canada with the possible exception of the Census of Population.

While CAI has now matured, the technology continues to evolve at a very rapid pace. In particular, new communications technologies and the Internet force statistical agencies to re-think their approach to data collection activities. In this context, this paper also discusses the current and potential applications of Electronic Data Reporting (EDR) in the Canadian setting. Few would debate that EDR is the way of the future. The critical questions that remain are how rapidly this method becomes acceptable to respondents and what tools and infrastructure should statistical agencies put in place to reap the maximum benefits from such opportunities.

2. THE EVOLUTION OF COLLECTION METHODS: AN OVERVIEW

This section will present an overview of the evolution of the collection methods used in survey-taking activities over the past 25 years at Statistics Canada. In the mid-nineties we have witnessed a significant evolution in data gathering methods . This is illustrated by Table 1 which shows the number of respondents per data-gathering method in the regions for establishment surveys since 1991.

TABLE 1

NUMBER OF RESPONDENTS PER DATA-GATHERING METHOD
REGIONAL OFFICES

SECTOR	YEAR	DATA-GATHERING METHOD				TOTAL
		MAIL	In-person interviews	Telephone interviews		
				#	%	
BUSINESSES	91/92	763,460	107,300	723,360	45.36	**1,594,120**
	92/93	651,130	96,640	784,270	51.19	**1,532,040**
	93/94	624,824	85,300	782,377	52.42	**1,492,501**
	94/95	409,405	85,720	771,293	60.90	**1,266,418**
	95/96	411,464	40,975	759,652	62.67	**1,212,091**
AGRICULTURE	91/92	-	12,700	153,230	92.34	**165,930**
	92/93	-	10,400	169,750	94.22	**180,150**
	93/94	-	24,380	207,708	89.49	**232,088**
	94/95	-	12,845	197,183	93.88	**210,028**
	95/96	-	12,848	199,974	93.96	**212,822**
PUBLIC INSTITUTIONS	91/92	37,980	140	27,550	41.95	**65,670**
	92/93	37,980	140	27,550	41.95	**65,670**
	93/94	36,100	252	28,500	43.94	**64,852**
	94/95	35,600	-	29,000	44.89	**64,600**
	95/96	2,338	-	12,000	83.69	**14,338**
TOTAL	**91/92**	**801,440**	**120,140**	**904,140**	**49.52**	**1,825,720**
	92/93	**689,110**	**107,180**	**981,570**	**55.21**	**1,777,860**
	93/94	**660,924**	**109,932**	**1,018,585**	**56.92**	**1,789,441**
	94/95	**445,005**	**98,565**	**997,476**	**64.73**	**1,541,046**
	95/96	**413,802**	**53,823**	**971,626**	**67.73**	**1,444,251**

The table brings several facts to light. There has been a significant decrease in the number of reporting units, a drop of approximately one-half million,mainly the result of considerable efforts in reducing response burden for business surveys. In particular, the Survey of Employment, Payrolls and Hours (SEPH) was revised in depth. A judicious use of tax data together with the creation of a smaller feeder business survey, the Monthly Business Payrolls Survey (BPS), significantly reduced the respondent burden and costs of this large scale survey. The table also shows a slight increase in the number of reporting units for agriculture surveys.

Secondly, mail questionnaires are no longer used for farm surveys, although that method is still preferred for business surveys, for reasons of economy. However, there has been a sharp drop in the use of this data-gathering method, i.e. close to 50%, which is twice the corresponding drop in the number of businesses.The use of the telephone to collect data in establishment surveys has increased overall from 50% to 68% over this five year period. Used extensively surveys (89% to 94%), this method has jumped from 45% to 63% for business surveys. In 1991, the ratio of reporting units responding by mail to reporting units responding by telephone was 1:1, and in 1996 rose to 1:2, a clear increase in the use of telephone interviews for business surveys. Let us look now at specific programs.

2.1 Agriculture Surveys

Until the early 1980s, the mail was the main method used to send questionnaires to farm establishments and gather the information to produce statistical indicators. Each year, approximately 200,000 questionnaires were mailed out and estimates were based on response rates of approximately 25%. Toward the 1970s, we started testing probabilistic methods in the agriculture sector, but it was only when surveys were revised following the 1981 Census that non-probabilistic methods were gradually replaced by probabilistic methods.

At the same time, because the sample was more targeted and smaller, we started doing telephone follow-up to maintain a representative sample. However, in the early 1990s, with the advent of computer-assisted telephone interviewing techniques, we began rethinking our data-gathering strategy and found that this method was particularly suited to these surveys.

2.2 Business surveys

For many years, many business surveys, used mail questionnaires for initial data gathering. This continues to be the case because it is still the most economic means of gathering statistical information for these types of units. On it's own, this method does not allow us to reach satisfactory response rates and telephone follow-up are required. Nevertheless (see Table 1), the percentage of units reporting by telephone has risen from 45% to 63% during the mid-nineties. There are several explanations for this increase.

- In the past, follow-up cards were mailed to encourage respondents to return the questionnaires by mail, but this produced very disappointing results and caused delays. This method was completely abandoned, and replaced by much more rapid and aggressive telephone follow-up. Occasionally, we follow-up by fax between the mail-out and the telephone follow-up. These follow-up methods are fully computerized and produce good results.

- Several monthly surveys are inputs to gross domestic product estimates. Target response rates for those surveys are generally 95%, and the time available for data gathering is short. The success of such an operation depends largely on good work organization and available resources. For these surveys, we analyse the response trends of each business and send a questionnaire by mail only to those units that respond faithfully with no telephone follow-up. In this way, businesses requiring regular telephone follow-up are simply deleted from the list of mail-outs and are contacted immediately by telephone at the beginning of the survey period, which helps balance the workload throughout the data-gathering period. For example, the telephone interview rate is approximately 50% for the Monthly Wholesale/Retail Trade Survey and 90% for the Monthly Survey of Manufacturing.

- As with farm surveys, we use CATI for business surveys, but mostly for telephone follow-up. The data from questionnaires returned by mail are entered separately. However, edit failures for those questionnaires are included in computer-assisted telephone follow-up, which allows us to benefit from the advantages of that method.

For special business surveys, we generally make a preliminary telephone contact to identify the individual within the firm who is capable of providing the information required, and to inform them about the survey and the importance of participating and filling in the questionnaire.

3. COMPUTER ASSISTED INTERVIEWING

Collection activities as Statistics Canada are conducted from regional offices and from headquarters. The increasing trend is for the regions to collect data for household based and social surveys while headquarters collection focuses upon business surveys. While the groundbreaking CAI pilot was an establishment survey extensive experience was gained in the regions with household and social surveys. Lessons learned in this context have contributed to the wider introduction of CAI to establishment and namely to business surveys.

Today, most survey collection processes at Statistics Canada use some form of Computer Assisted Interviewing (CAI). By contrast, in 1987, Statistics Canada had one on-going survey in production using CAI. The ten years that followed were ones of tremendous change. This section reviews the CAI transformation at Statistics Canada and highlights some of the major issues encountered in the process.

3.1 Early Developments

Computer Assisted Telephone Interviewing (CATI) as a technology was well established when Statistics Canada began development in 1985 of its first project.

The survey involved was the Survey of Shipments, Inventories, and Orders (SIO), now referred to as the Monthly Survey of Manufacturing (MSM). This survey employed a mail-out/mail-back methodology with telephone follow-up for non-response and edit failures. Additionally, there was a small subset collected over the

telephone. This subset was made up of units with a historically poor response or which were completed over the telephone at the respondent's request.

This initial project did not utilize computer assisted interviewing software. It was developed using proprietary software called Data Entry Facility II (DEF II) that ran on a Honeywell DPS 6 mini-computer. The entire system - the CAI instruments, the call scheduler, and interactive edits - were all custom coded.

The motivation for developing this survey, as a CAI project was primarily to 'break the ground' and generally become familiar with computer assisted interviewing. There were expected minor cost savings associated with clerical operations and data entry. The real significance of the project was in the lessons learned. A series of management reports had been designed with the initial implementation of the survey. Over time, these were refined and formed the basis for a standardized set of management reports we expect from a CATI system. Questions about the training of interviewing staff were also answered. There had been an expectation that more experienced interviewers might have difficulty adjusting to an increase in technological sophistication and change in methods. This was not the case but the project did re-emphasize the importance of in depth training of all involved staff and a phased in conversion to CAI. The switchover was in fact quite smooth, due in large part to an extensive interviewer-training program.

As the conversion of SIO to CATI was being phased in, a major study of the effects of CATI on costs and data quality was being developed. This second test was developed using the United States Bureau of the Census (USBC) 'Mini-CASES' software running on a VAX mini computer. 'Mini-CASES' is a variation of the University of California at Berkeley Computer Assisted Survey Execution System (CASES). This test compared results collected in a centralized interviewing facility using paper and pencil methods and CATI. The study concluded that quality gains would likely be the most important impact of CAI. It also suggested that there might be a marginal increase in costs because of CAI. This was due to slightly longer interview times.

These two early CAI projects generated considerable discussion within Statistics Canada and established the climate within which further CAI development became possible.

3.2 Computer Assisted Telephone Interviewing

Telephone surveys and follow-ups conducted from the head office or regional offices, provide an environment conducive to tighter work and quality management. In particular, the CATI method, used in almost all telephone surveys for approximately eight years, has several advantages, including interactive monitoring, automatic routing of questions, and above all, a highly effective method of assigning calls to interviewers and managing calls. All these factors help reduce costs to a minimum. Furthermore, the monitoring of interviews for which a structured and formal methodology was developped at Statistics Canada significantly enhances the quality of the interviews and of the data-gathering tool .

3.2.1 The use of CATI in agriculture surveys

CATI has proven to be most effective for the conduct of agriculture surveys. As Table 2 shows, farm surveys require data from a fairly high number of operators over a very short period of time. Because the length of the interviews is fairly short (8 to 12 minutes) and publication deadlines very tight, the CATI method was particularly well suited for these applications. Very rapid and effective, the survey process has been greatly simplified and improved by the introduction of interactive controls.

3.2.2 The use of CATI in business surveys

The use of CATI for business surveys represents a major thrust in the introduction of CAI. These mail-out/mail-back survey programs made extensive use of telephone follow-up for non- response and edit failure. The work was characterized by the involvement of a multiplicity of subject matter divisions with a variety of approaches to data collection. Consequently, the conversion to CAI has been slower than the transformation in social surveys where there was basically only one sponsoring division. Overall survey re-engineering has often marked the conversion to CAI in business surveys.

TABLE 2

CATI FARM SURVEYS

Some 1995 examples

Survey	Size of Sample	Average length of interview (minutes	Data-gathering period (days)	Number of interviewers	Response Rate
Survey of Fruit and Vegetable Productio	20,000	8	16	43	97%
June Farm Survey	28,000	9	10	99	90%
November Farm Survey	27,000	10	10	73	93%
July Livestock Survey	27,000	8	12	81	97%
Survey of Greenhouse and Nurseries	3,200	22	17	17	94%

The typical survey today employs a mail-out/mail-back questionnaire that is tailored to the individual respondent based on pre-determined commodity classifications. Follow-up of non-respondents and failed edit cases are done in CATI mode. The call scheduler within the CATI software controls all aspects of the collection process. This is an innovative, complex and highly integrated approach to data collection. It is a significant change from the pre-CAI approach. The annual cost savings are approximately $600,000, including 20 person-years of work.

This broad-based move to CAI methods has impacted on the organization. The change from paper based systems to automated systems has resulted in a tremendous growth in informatics support within each region and head office. At the same time, the move to automated systems has required a significant upgrading of the skill set of project supervisors and program managers. The new technology has brought some significant improvements in efficiency and effectiveness. Conversely there has been some of turmoil as staff adapts to a climate of accelerated and continuous change.

The technology continues to evolve. Audio and video capabilities are beginning to be explored. New opportunities to interface directly with businesses' own information systems are being investigated. We have also taken the strategic decision of moving all CAI applications to single software (BLAISE) and hardware platforms. This will introduce flexibility in moving workloads between the field and office operations as well as between the regions and headquarters

3.3 Mixed Mode; Computer Assisted Personal Interviewing

Computer Assisted Personal Interviewing (CAPI) typically involves the use of portable computers by field interviewers to conduct face to face interviews. It can, of course, also be used for telephone interviewing conducted out of interviewers' homes. It is worth noting that CAPI differs from CATI in as much as the latter are office operations whereby interviewers work at work stations, networked and linked to a server which manages the scheduling and allocation of work.

The conduct of a mixed mode collection (CAPI, EDR/INTERNET, CATI) for the Workplace and Employee survey was a significant accomplishment in the establishment survey field

This survey is a longitudinal annual survey of both employers and employees. This survey collects data from a sample of employers and employees on workforce characteristics such as: job organization, compensation, training, human resource practices, collective bargaining, workplace performance, business strategy, innovation, technology use and use of government programs.

For the odd year's (i.e.1999), both CAPI and CATI are used to conduct the data collection. For the even year's (i.e. 2000), all of the data collection is done using CATI. In the odd years, the sample of employers is strengthened to replace units that were out of business etc. and the entire sample of employees is refreshed.

Prior to the start of the first year, advance research was conducted on all small businesses to replace units in the sample that were out of business and out of scope. In addition, due to cost constraints the project teams decided to conduct a portion of the employer sample (about 3000 employers) full CATI. In addition to completing the employer questionnaire CATI, Interviewers obtained a list of employees to select employees for the employee survey. Selected employees were mailed an introductory package via their employer and the employee sample was forwarded to the collection site conducting the employee survey. These were small employers that had less than 10 employees. The employer CATI operation was conducted out of the Atlantic region.

The remainder of the sample (about 5000) employers was done CAPI.

All of the CAPI/CATI collection systems were developed using the CASES software. The survey file was sent out and loaded on a Business Regional database. This base permitted us to print address labels and print tailored employer questionnaires for secondary contacts. The database was updated from the pre-contact module. What follows is a brief description of each module.

The *pre-contact module* verified the business and contact information. Businesses that were determined to be out of scope or out of business were coded out and were transmitted back to head office via the employer module. All valid units were assigned for a CAPI interview. Units were assigned on a daily basis as soon as the laptops were available. For businesses with multi locations, reporting arrangements were negotiated in advance of the CAPI interview. Respondents were given the option of getting more than one contact involved in completing sections of the employer questionnaire. If secondary contacts were identified they were mailed a tailored questionnaire with only the section(s) they were to complete. Section A of the questionnaire was always sent to the primary contact for a CAPI interview.

A separate communication system was devised using the Internet for interviewers to pick up assigned units and send back completed or partially completed units back. This system was centrally controlled from head office. Out going units were loaded on a WEB Site. In coming units were encrypted and passed through an approved transmission process. The pre-contact operation was responsible for monitoring and following up on any outstanding sections of the employer questionnaire for secondary contacts.

The *employer module*, CAPI based, was conducted in the five regions: Atlantic, Quebec, Ontario, Prairies and Pacific. Local interviewers conducted the CAPI interviews in the major centres. For remote locations or for locations where it did not warrant hiring an interviewer to conduct a CAPI interview, the units were assigned to a laptop in the regional office. The interviewers then completed the employer questionnaire by telephone and made arrangements for an interviewer to visit to select a sample of employees.

Units that were coded out on the pre-contact module were outputted to the employer module and then were transmitted back to head office. Units that were returned from the laptops (CAPI) were edited for any missing information. The employee sample was retrieved and sent on to the collection site conducting the employee survey CATI. Unresolved edits were followed up by telephone.

The *employee module* was conducted out of two sites (Winnipeg and Sturgeon Falls). The employee sample was received as soon as the employer unit was returned from the laptop or was completed in the case of the employer portion that was done 100% CATI. Staff waited two weeks to receive the employee form that was mailed out to selected employees in their introductory package. This form was used to solicit the participation of employees in the survey. The form also collected the data for four questions that were mandatory for the employee survey.

Some valuable lessons where learned from the first year application.

- More time and volume testing is needed to make sure that all systems are functioning correctly in an integrated fashion. The individual laptops that are set up for an individual interviewer must be thoroughly tested to ensure passwords are working. The laptop programs that decrypted incoming files and moved them into CASES and encrypted and sent completed units back were not robust enough. Several patches had to be sent out to try to fix this problem. As soon as the laptops are in the hands of the interviewers it is very difficult to fix a problem. There were also problems with incorrect security passwords being wrongly assigned to laptops.

- The communication systems, although sound, were not well understood or supported by regional technical staff. More in depth and formalized training is required

4. ELECTRONIC DATA COLLECTION

Recent years have witnessed the wide scale adoption of personal computers in the workplace and in households. The increasing user friendliness of operating environments and software suites has broken down cultural resistance by facilitating their use. The World Wide Web through the Internet and associated developments has expanded our opportunities to collect data and possibly reduce response burden.

The advent of electronic data reporting (EDR) can be viewed as a natural extension of the use of technology in data collection activities. Whereas CAI effectively combined into one step operations that were distinct previously, such as interviewing, data capture and editing, EDR goes a step further by shifting such activities to the respondent.

It should be stressed that the expectation is not that EDR will replace other collection methods in the near to medium future. Our objective has not been to replace one by the other, but rather to develop the best set of collection tools for any given survey, that will make the task of the respondent as easy as possible, by providing as many options as possible. More generally, as a statistical agency, we consider it our responsibility to be continually looking for ways to facilitate the task of responding. This hybrid approach has been highly adaptable and flexible for both survey respondents and survey managers.

The development of an electronic data reporting capability has presented us with numerous opportunities and challenges.

4.1 Opportunities and Challenges

The capability to offer our respondents collection options that reflect the state of today's communication technologies facilitates participation in our surveys and can strengthen our relationship with respondents. The potential for applying edit verifications in the respondent's environment can improve quality at source, which can eliminate the need for subsequent follow-ups consequently reducing the response burden. The use of electronic means of delivery and retrieval of survey questionnaires can improve general timeliness.

The transfer of the data capture and editing functions to the respondent, combined with improved data quality at source, greater timeliness in data collection and reduced follow-ups, can lower overall survey taking costs.

Many of the challenges that the electronic data collection method presents us with have either been met or are currently being addressed. We must focus on greater value added to the respondent and intuitive processes to ensure a maximum take up, if we are to recoup our investment.

In this same vein, we must define and carry out the appropriate marketing and communication strategies. The most significant challenge lies in providing an infrastructure, methods and tools that protect data confidentiality and security, and answer the respondent's concerns. The challenge is even greater in ensuring data security and confidentiality in a real time, on-line environment, in a cost-effective fashion. We must also consider when applying edits in the respondent's environment, what constitutes the proper balance between value and hindrance. The retrieved data must also be seamlessly reintegrated into existing survey databases, without undertaking costly system redesigns.

4.2 Evolution

While we tend to perceive electronic data collection as a relatively recent development some form of it has being going on for quite some time. For the last decade we have been collecting data from other government departments and institutions on tapes and cartridges.

Over the last ten years we have collected, from a score of financial institutions, data files using a modem to modem connection. Pre-arrangements for transmission are made with the institutions. The file transfer is monitored on a public network workstation and once complete is copied to diskette, then deleted from the hard disk. The diskette is then moved to our secure network. The respondents are notified by telephone of a successful or unsuccessful transfer.

More recently, we have been collecting more than three hundred establishments of our Annual Retail Trade Chain Store Survey using a diskette-based system that we refer to as the Personalized Electronic Reporting Questionnaire System (PERQS). The application, which is FoxPro based, applies a complete set of edits, historical checks and has the ability to import data into the survey form from an Excel or Lotus spreadsheet. The application is loaded in the respondent's environment from a diskette and the completed survey form is also returned by diskette. The content of the diskette is encrypted and courier to and from the respondent.

4.3 More Recent Developments

In the last three years, an important step forward was the putting in place of a corporate-wide infrastructure to transfer data to Statistics Canada over the Internet. The infrastructure consists in FTP and SMTP public and secure servers with robotic switching between the two sectors.

Security has been addressed through encryption algorithms applied to the source data. Integrated applications installed on the respondent's workstation allow him/her to encrypt and return data files or survey forms with capture and edit capabilities in multiple modes. The applications are distributed through a download from a Statistics Canada's, Web site or from a CD-ROM. The survey forms are deployed in a Windows point and click environment. Once the data are received and in a secure environment, they are virus checked, decrypted, verified for integrity and recovered if need be, verified for duplication, and reintegrated into survey databases or pushed to client divisions. Reports are generated to monitor the process.

This infrastructure has now been used successfully in a number of more recent projects. Universities, Government departments and Customs use this infrastructure and applications to report administrative data through electronic files to our Education, Transportation and Financial Statistics programs. Canadian Exporters and respondents to the Unified Enterprises Surveys also use the same tools and environment to return completed survey forms to our International Trade and Business Surveys programs.

Respondents to the Business Payroll Survey (BPS) do the same using a structured data file based on Statistics Canada specifications. This structured file can then be wrapped in an Electronic Data Interchange (EDI) envelope and be returned through a value-added network.

The application developed for the Unified Enterprise Statistics Surveys is significant in it's capability to maintain on a common platform up to twenty one distinct surveys in both of the official languages, with the Wholesale Trade Survey form capable of being customized at the commodity section level by the respondent.

The Canadian Exporters application has a commodity coding assist and auto coding capabilities, as well as a mapping assistant function, to relate source file elements to survey form elements. Once this relationship is established a data import to the form is enabled, eliminating the exporter's data capture cost.

4.4 Current Research and Development

The thrust of the current research and development activities lies in the development of a centralized collection Web site at Statistics Canada. The site will provide easy access to relevant information about the survey.

The survey forms, edits, concept and definitions, help, encryption, send via electronic mail will be deployed from a Statistics Canada public Web site in a browser environment. Depending on the survey, coding assists and auto-coding functions will also be available. The respondent will be capable of completing the survey form or forms in one or multiple sessions using a unique identifier based on questionnaire identification provided by mail-out. The security and confidentiality of the data will be ensured by creating a database on the respondent'

workstation and not on the Web server. It is the local database that will be encrypted and sent via a separate route to Statistics Canada's public FTP or SMTP servers.

A series of generic functions such as a mapping assistant for data import will be made available. Import from commercially available software products will also be possible. To this effect, Statistics Canada will make available, from this Web site to software developers, its data definition and record layout specifications. In parallel, consultations with major electronic filing software developers are in progress, to pilot this approach for the Quarterly Financial Survey. The ability to import into the electronic survey form will certify the product as being compliant with Statistics Canada requirements.

The Internet Service Providers Survey and the current users of the redesigned Personalized Electronic Reporting Questionnaire System (PERQS), mentioned previously, will be the first to avail themselves of the Web based collection facilities, and implementation is scheduled this fiscal year. It will still be possible to deploy the entire application on a CD-ROM for respondents who indicate this preference.

After the implementation of the Web based approach we will expand this option to more surveys. We will use the first year experience to fine-tune the product and the environment. We are conducting research, for less complex surveys, to identify mechanisms whereby we could collect data directly on the Web and store the data on the Web server as it is being entered. We will seek solutions to encompass more operating platforms. We are also conducting research to define the appropriate Public Key Infrastructure to address specific requirements for the application of historical edits.

5. CONCLUSION

In this paper, we have attempted to describe the evolution of collection methods at Statistics Canada. Considerations of efficiency and effectiveness have lead to an increased use of telephone interviewing. Telephone interviewing has therefore become an important tool in Statistics Canada's survey program.

In the short and medium terms, we do not foresee a decline in the use of telephone interviews. The advantages offered by this method in terms of flexibility and rapidity make it the preferred choice for anyone seeking to balance costs, quality and deadlines. There are however some clouds on the horizon. The proliferation of cellular phones may compromise the use of random digit dialing in the medium term. The widespread use of CAI information gathering, whether for a government statistical agencies or polling firms, may begin to produce a saturation effect in the population, reducing the take up of this method by respondents. We will need to monitor closely and adapt our strategies accordingly

Computer Assisted Interviewing has now become a standard way of doing business. Not only has this lead to significant savings in our collection costs, it has created opportunities to undertake surveys that are far more complex than ever before. In our context, interviewers have adapted better than expected to this new environment.

However, the change to this new and highly complex environment has established the need for higher and more competent levels of technical resources. It has also challenged our project teams responsible for survey planning and testing; in as much as the requirements of CAI environment are far different from the traditional paper and pencil.

One of our major priorities has been and continues to be the facilitation of participation by respondents to our surveys. One of the means of achieving this objective has been the diversification in methods available to the respondents in providing the department with the needed statistical information. It is in this context that pioneering work on electronic data reporting (EDR) making use of the Internet was initiated.

In the longer term, EDR has the potential to become the primary data collection method. Nevertheless we will continue to offer a multiplicity of reporting means to reflect our respondents diversity. Greater confidence in the new medium particularly in regards to security and confidentiality will gradually yield increased take-up. Increased technical readiness in the respondent's environment will also be contributory. The self resolution of bandwith issues over time will also spur growth of the EDR method. Notwithstanding, it is imperative that statistical agencies such

as Statistics Canada actively pursue research activities to perfect this method and any other that has potential for the respondents thus stimulating participation .

REFERENCES

Gosselin, J. -F., Williams, B.J., Developments in Data Collection at Statistics Canada, The International Symposium on New Techniques of Statistical Data Acquisition, Tokyo, Japan, 1999

Gosselin, J. -F., La pratique des enquêtes par téléphone à Statistique Canada, 5Ièmes journées de méthodologie statistique-INSEE, Paris, France, décembre 1996.

Catlin, G & Ingram, S. "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing" in Telephone Survey Methodology; Groves, R.M. et all editors; John Wiles & Sons, New York, 1998.

Mudryk, W., Burgess, M.J., Xiao, P., A Quality Control Approach, to CATI Operations in Statistics Canada.

Ménard, M., Parent, G., " Electronic Data Collection" in Symposium 97on New Directions in Surveys and Censuses, Statistics Canada, Ottawa, Canada, 1997

PROGRESS AND PROJECTIONS IN COMPUTER ASSISTED DATA COLLECTION AT THE BUREAU OF LABOR STATISTICS

Richard L. Clayton, Richard J. Rosen, William McCarthy and Jim Kennedy
Richard L. Clayton Suite 4840, 2 Massachusetts Ave. NE, Washington DC 20212 Clayton_R@BLS.gov
Bureau of Labor Statistics

ABSTRACT

During the summer of 1994, an internal BLS survey was conducted on the status of automated data collection in the Bureau. The survey established a baseline for evaluating the status and progress of BLS programs in identifying and exploiting methodological opportunities to improve data quality. Also, that report on the Computer-Assisted Survey Information Collection (CASIC) survey and its results provided a means for sharing information across BLS programs on CASIC initiatives. The survey also provided a 5-year projection on methods expected to be employed and sample proportions to be collected via CASIC by the year 2000. This paper reviews the current status of CASIC implementation, compares it to the earlier projections, and looks forward to the next five years. Lastly, this paper describes factors leading to and restricting CASIC implementation.

Keywords: CASIC, Internet, Electronic collection, EDI

1. INTRODUCTION

The 1980s brought about many dramatic technological changes in the business workplace, most notably the widespread use of microcomputers, telecommunications, and electronic information exchange. The availability of these new technologies offered statistical agencies many new opportunities for improving the timeliness and quality of data collection for establishment surveys. Traditional mail collection has always severely limited the timeliness of the published estimates from establishment surveys and, thus, their usefulness for current economic analysis. In contrast, the opportunities offered by new automated collection approaches afforded almost instantaneous access to employer data. Some of the automation options even offered improved timeliness and control at costs lower than traditional mail collection.

The term CASIC was coined in the 1990 Statistical Policy Working Paper 19 entitled "Computer-Assisted Survey Information Collection" under the auspices of the Federal Committee on Statistical Methodology (FCSM). CASIC refers to the growing number of data collection methodologies employing the rapidly advancing set of technological tools available through inexpensive microcomputers and laptops. CASIC includes such methods as CATI, CAPI, TDE, VR, CADE, FAX, CSAQ, Internet/Web-base, and EDI, each of which is defined briefly in the next section. The development of a wide range of CASIC methods represents the most significant advance in survey methods in the last two decades.

The last 15 years has seen an explosion of research and number of CASIC methods, and the Bureau of Labor Statistics has been in the forefront through testing and implementing existing CATI and CAPI methods, through the first-time development of new innovations such as Touchtone Data Entry and Voice Recognition, Internet/Web-based and Electronic Data Interchange. The survey yielded a profile of current CASIC uses and plans for the remaining years of the 1990s.

2. BLS CASIC USES

By 1995, each of the three large survey offices in BLS used at least one form of CASIC in one or more surveys. Overall, a large variety of CASIC methods currently are employed in BLS. The Current Employment Statistics (CES) survey has devoted 15 years to developing and implementing a wide range of CASIC methods including CATI, TDE, VR, Web-based, and EDI for data collection and CATI and auto-FAX for nonresponse prompting. The data for several programs were collected totally by CASIC in 1995 including the CPS, NLS, and SEPT.

3. PLANS FOR EXPANDING CASIC METHODS

During 1995, program managers were asked to predict their CASIC status over the next 5 years, to the year 2000. Their responses indicated that, by the end of the decade, virtually every program would use one or more CASIC methods. Indeed, the next five years were to see an acceleration of CASIC implementation, including:

- the CES will complete CASIC implementation around several CASIC methods, including use of the Internet for collection;
- the ES-202 and OES programs will use EDI;
- SIC refiling may be conducted via TDE for most respondents;
- the CPI will complete full-scale CAPI use;
- the PPI will implement CAPI for solicitation and test and evaluate TDE and EDI for repricing; and
- OCSP will use EDI in several programs.

The concept and scope of CASIC is usually interpreted as being limited to data collection. This notion is being re-thought. The same technologies which represent the forefront of CASIC, such as Internet, and the same methodological issues we have faced in our CASIC development now offer rich opportunities for improving other survey operations such as editing and information dissemination.

4. CASIC METHODS DEFINED

The new microcomputer-based collection methods have altered survey agencies' approaches to both interviewer-collected data and mail-collected data. A summary description of each method is provided.

Computer-Assisted Telephone Interviewing (CATI) offers a structured approach to traditional telephone collection. Under CATI, the survey questionnaire, along with full editing capabilities, are resident on the computer. The CATI interviewer accesses the next scheduled sample case and autodials the respondent. All scheduling of cases, along with callbacks, are usually controlled via a computer scheduling algorithm. The interviewer takes the respondent through a computer controlled sequence of questions, resolves all edit failures, and then schedules any subsequent recontact which may be required with the respondent.

Computer-Assisted Personal Interviewing (CAPI) provides a structured approach to personal visit collection. In this case, the questionnaire and editing capabilities are resident on a portable computer carried by interviewers in the field. All other interview functions are similar to CATI and the data are normally electronically transmitted from the field site to the agency's mainframe on a regular basis.

Touchtone Data Entry (TDE) offers an alternative to mail collection for a number of applications. Under TDE, the respondent uses a touchtone telephone to call a toll-free number which is connected to the agency's touchtone computer systems; this activates an interview session. Again, the questionnaire is resident on the computer in the form of prerecorded phrases and each question is read to the respondent. The respondent uses the touchtone telephone buttons to enter the numeric data requested. Each answer is read back for respondent verification. Limited editing is also possible under this method; however, survey agencies currently are relying on interviewer followback calls for edit reconciliation. TDE eliminates mail handling activities, mail delays, and key entry activities.

Voice Recognition (VR) is a variation of the TDE collection process whereby the respondent only needs access to a telephone. The CES VR system is speaker independent and accepts continuous speech; it recognizes digits zero through nine and "yes" and "no." The respondent's verbal answer is translated by the VR system, then the answer is repeated by the VR system for respondent verification.

Internet/Web-based collection holds tremendous promise for the most radical change in establishment collection of all CASIC methods. The user-friendly interface will be exploited to provide data entry and editing, as well as other facilities such as easily updating respondent and establishment information. Most significantly, Internet/Web-based collection offers major cost-reducing features beyond TDE (where editing is not easily done) and will reduce the need for interviewers as seen under CATI. The role of interviewers will likely evolve into more narrowly targeted areas such

as reluctant respondent conversion and non-response prompting. Security concerns have slowed adoption of Internet collection, but new techniques are being readied to allow for its use. The natural advantages of this method will undoubtedly lead to its widespread use in the next few years.

Computer-Assisted Data Entry (CADE), much like CAPI in principle, allows on-line post-collection entry, editing, and coding for improving data quality.

Electronic Data Interchange (EDI) offers the ability to collect large volumes of data from businesses. Respondents extract the needed data in a pre-specified format from their computer databases and initiate electronic transmissions to the survey collection agency. Basic EDI systems simply accept data files (i.e. respondent recontacts for edit reconciliation are done later by interviewers); more sophisticated EDI systems can offer direct on-line editing by the respondent and a number of automated messages.

Computerized Self-Administered Questionnaire (CSAQ) entails providing software to respondents with access to a microcomputer. It works much like CATI except that the respondent interacts directly with the software on their own computer. The software prompts for answers, branching as needed and conducts on-line editing. When completed, the respondent may either mail the diskette back or, if they have a modem, initiate a "send" function transmitting the completed data file back to the agency. This method will very likely become one of the fastest growing methods as several agencies begin examining how to exploit the Internet; the most oft-cited approaches entail electronically sending CSAQ instruments. A broader definition of CSAQ includes TDE and VR.

Each of these new collection methods has several common advantages. They represent an evolution towards a paper-less collection environment whereby all questionnaires and reported data reside only in electronic form, usually vastly reducing or eliminating labor-intensive mail handling activities.

It is very likely that a single CASIC method will not address all of the reporting situations encountered in establishment surveys leading. This variety in reporting arrangements may lead to a mixed-mode environment depending on access to technology by respondents and also on the volume of data provided. Each CASIC method also requires that a level of technology be available to the respondent. While TDE assumes a touchtone telephone, VR requires only a telephone. EDI requires that the respondent have access to a central database and be willing to invest in extract programs and use on-site technology to initiate and send files electronically.

Full-scale implementation of CASIC in the CES focuses primarily on use of TDE and VR with smaller portions of the sample collected by other methods. Over 90% of respondents have touchtone telephones and report for 1 to 6 units making them suitable for TDE. Respondents providing data for large numbers of units and access to central databases and telecommunications technology may report via EDI for inexpensive, low-burden reporting. For respondents reporting for more than six locations, where TDE and VR may provide very long interviews, specially-designed forms may be faxed with automatic character recognition. It is likely that a small portion of respondents, only 1 to 5 percent, may report via mail.

Selecting CASIC methods should be done with clearly defined problems and knowledge of the sample characteristics and technological requirements placed on the respondents. Although it may be possible to offer respondents a range of reporting methods as a means of obtaining their participation, each method varies in cost and timeliness factors, as well as suitability for solving particular survey problems.

5. RE-ENGINEERING

CASIC methods are improving data quality, timeliness, and reducing respondent burden and costs. The widespread use of CASIC is re-engineering the survey data collection process and will continue to cause ripples of change in other survey functions.

The effects of CASIC are being seen in the composition and characteristics of our survey organizations. For example, when data are collected electronically, staff devoted to mail handling and key entry tasks can be replaced with a single

TDE desktop computer backed by a small programming staff, changing the cost structure from labor-intensive to capital-intensive. Ever-increasing postage costs can be replaced with lower and declining costs of telephones.

CASIC also acts to move many post-collection activities to the immediate point of data collection. This work can be most effective in time-critical surveys by eliminating or minimizing time-consuming and expensive review. For example, CATI and CAPI build in on-line edits resulting in clean data that need no subsequent review. EDI methods include prior validation of data files insuring data quality prior to submission.

CASIC systems also can be built to capture process control information allowing careful quantitative analysis of each step as a basis for continuous improvement. CASIC can provide the means for introducing changes not possible under previous methods. The best example here is the new CPS questionnaire implemented under CAPI which is too complicated for paper and pencil interviewing.

Implementing CASIC may mean merely automating existing manual procedures with important, but limited, improvements. However, the implementation of a new methodology also offers an opportunity to do much more, to achieve a true "breakthrough" in the overall performance standard. To fully capitalize on such opportunities CASIC methods may well involve changes to forms, the content and timing of operations, the entire nature of the respondent contact program, and the organization and staffing of field operations. Thus, large-scale CASIC implementation truly represents a re-engineering of the data collection process, with resulting improvements in quality, timeliness, and changes in cost structure.

6. CASIC SURVEY AND RESULTS

A profile of current and projected CASIC uses by survey function is shown in Figure A. Overall, a large variety of CASIC methods are currently employed in BLS.

CES: The CES survey has focused a decade of research and development on a number of CASIC methods toward accelerating data collection to reduce the size of revisions at the lowest cost. The CES uses several methods for data collection including CATI, TDE, VR, and most recently EDI. In addition, CATI and automated outbound FAX are used for nonresponse prompting in an effort to completely automate data collection.

ES-202: EDI is used for large multi-unit firms. Also, SIC refiling has taken an interim step by using barcoding of "No Change" forms.

CPS: The CPS is completely CATI and CAPI collected. Both the NLS-Y and NLS-W surveys are collected completely by CAPI.

CPI: The Consumer Price Index began investigating CASIC methods in the late 1980s. Most early research and development focused on centralized CATI collection, and a test of a CATI instrument for the Housing component began in early 1990. A test of CATI for Commodities and Services (C&S) followed the next year.

OCWC: In three major surveys, CADE is being used as the first step in improving Occupational Compensation Survey Program (OCSP). Also, a prototype expert system was designed for job matching during data collection. This current system could be used to help field economists write and file interview results and train new staff in job matching.

OPT: The Hours at Work survey uses CATI for non-response prompting, including collecting and editing responses not received by mail.

OPUBSS: The SEPT I survey uses a Windows-based system incorporating both CADE and CATI functions. Data collectors use CADE with on-line edits for data entry of mail or faxed questionnaire surveys. If an establishment questionnaire contains a large number of edit errors, the respondent is called back for edit reconciliation using CATI. CATI is also used to collect and edit data from mail survey nonrespondents. Nonrespondents to the mail survey also can receive a data collection instrument via automated fax.

The SEPT 2 survey uses a similar system incorporating both CAPI and CADE functions. Field economists using a laptop computer can collect data directly from respondents using CAPI or can enter the data subsequently by using CADE.

7. ACCELERATING CASIC USE AT BLS

CASIC began in the 1970's with CATI using mainframe and mini-computers, but applications exploded with the availability of PCs and then laptops. Figure X shows an accelerating pace of tests and implementation in an increasing number of surveys since 1990. By the end of the decade, virtually every program plans to have moved a substantial portion of collection to one or more CASIC method.

CES: In addition to the CASIC methods currently used, the CES is testing outbound FAX as a means for editing data by the respondents. Further, the CES is beginning tests using the Internet/E-mail for collecting data. SIC refiling now involves contacts with up to 2.2 million businesses each year. Based on the CES TDE system, tests involving TDE are now studying whether the estimated 66 percent of respondents which report that their SIC and other information did not change from the previous period would simply call a toll-free number and indicate "No Change" by pushing a few buttons on the telephone. If successful, postage, mail handling, re-mailing, and review workload could be vastly reduced for those units.

OES: The Occupational Employment Statistics Survey plans testing of EDI. Tapes have been received from selected large respondents for years.

MWR: Multiple Worksite Report program also plans to use EDI to receive data. A standard data file format has been developed.

IPP and PPI: The International Producer Price Index (IPP) and Producer Price Index (PPI) are developing a computer-assisted initiation system to: 1) support coordinated initiation between the two surveys; 2) reduce the resampling burden for PPI overlap units; 3) move data capture and edits closer to collection; 4) support electronic data transmission; and 5) provide more flexibility in collection materials. An operating strategy and plan were completed in 1995, and a prototype initiation instrument is being developed for CAPI using pen computers. The final CASIC overview of the PPI could include CAPI and CSAQ for initiation and EDI, FAX, E-mail and/or TDE for monthly repricing.

Wages: EDI will play a major role in ECI, EBS, and OCSP over the next four years, particularly for large, multi-unit respondents. Also, the OCSP Reinterview Program will be completely converted to CATI and CAPI.

8. 1994-1999 PROJECTIONS

In general, significant advances were made in CASIC implementation, as shown in Table A. The CES continued its drive towards full implementation with now – percent of the 390,000 units collected by CASIC, with 250,000 using TDE each month.

However, few programs made the rather ambitious projections. There are several reasons for this relative shortfall. Primarily, most things look “doable “ in a 5-year period. In reality, many new program objectives can interfere with an internally driven restructuring of operations. For example, the CES is now in the midst of a full-scale redesign to a probability sample and other critical program changes, and the ES-202 will move forward using TDE for the Annual Refiling Survey once the transition to the new North American Industry Classification System (NAICS) is completed. Also, projections may have been based on expected budget changes, which may not have been realized.

9. 2000-2005 PROJECTIONS: WHERE WILL WE BE?

9.1 The Next 5 Years

Over the next five years, BLS programs again are projecting significant shifts to CASIC methods. While the more traditional CASIC methods will see additional usage, Internet/Web-based collection will see the greatest increase. The CES began using Internet collection in 1996 on a limited, research basis and has retained the sample use at small levels. Security concerns and other program priorities kept this number small, however, it was an important learning experience. BLS now feels that security of data can be guaranteed and several programs are now making and expect to continue making, major advances including expanding use in the CES and implementing WWW based collection in the Producer Price Index, the Multiple Worksite Report (MRW) and Annual Refiling Survey (ARS) in the ES-202 program, OSH and OES.

Virtually every program will use CASIC in some manner by the year 2005. The benefits of CASIC will be essential to continuing the Bureau's role as a leader in providing highly accurate and timely economic data. The continuing challenge to do more with less also argues for continued investment in quality-enhancing, cost-reducing methods.

Importantly, given the increasing use of CASIC methods, the Bureau is well-positioned, with a growing number of experienced staff, to meet old and new challenges. The availability of experienced staff lessens the potential for re-inventing the wheel and for sharing hard-learned lessons.

9.2 The Future Of Mail Collection:

However, this viewpoint should be placed in context. BLS identified 14 establishment-based data collection programs. These programs vary widely both in terms of periodicity and sample size. The combined sample for these programs is over 5 million units with over 13 million annual responses. Some of the programs are relatively large, such as the Quarterly EDI (6,600), while some are very large such as the Annual SIC Refiling Survey (ARS) with over 2.7 million per year.

In reviewing these modes a couple of general conclusions can be reached:

- For most programs, non-CASIC collection still predominates as the primary collection mode;
- Mail is still the predominant mode of collection for most programs (5 of the 14 programs);
- CASIC methods are primary in only 5 of the 14 programs.

For all of the CASIC work, mail-based collection remains the primary mode in many establishment surveys with CASIC playing a secondary role. A couple of other, not as obvious, conclusions can be made. Most of the very large data collection efforts continue to be mail-based, whereas, smaller surveys employ CASIC methods. The only large survey to employ CASIC methods as a primary mode of collection is CES, where touchtone data entry collects over 60% of the responses. This indicates that it may be easier to implement CASIC on smaller samples.

One may ask: Why has the Bureau not moved more rapidly into the CASIC sphere? What are the roadblocks to CASIC implementation? Is this simply resistance to change, or are other factors involved? Can these be overcome?

There are a number of factors to consider in this area. Some of the major factors are described below:

Funding/Cost: Moving forward with new technology often requires an initial outlay of resources. Sometimes these resources cannot be adequately justified either within the agency or to others.

Technology: Sometimes what appears to be a promising new technology often, after testing and prototyping, fails to meet expectations.

Management: Sponsors of new technology need to be able to make their case to upper management in order to obtain the necessary resources for implementation. In some instances, we may not have done as good a job of selling the methodology to management.

Human resource constraints: There is a growing need both in the government and the private sector for talented IT professionals. May feel that the government agencies have been hampered by in their IT efforts in general by their inability to offer attractive wage and benefit packages offered in the private sector. This lack of cutting-edge IT staff may permeate all efforts at CASIC techniques.

Data security/confidentiality: We are all becoming increasing aware of the problems and issues related to the security of the Internet. However, other security issue must also be considered in any CASIC implementation. Most CASIC methods eventually involve using computer, networks and other infrastructure that may be more susceptible to intrusion. BLS has instituted a special clearance/approval procedure for all current and future electronic collection methods. All current/future electronic collection methods must be reviewed by a Security Committee that will review the hardware, software, for compliance with BLS standards and the data security procedures in place.

Staff reluctance/Training: While there is little evidence that this is a factor, "reluctance to change" can often permeate an organization. This often does not directly manifest itself, but may show up as other road blocks such as "I don't have time" or "I don't have training."

10. FUTURE CASIC METHODS

Survey methodologists are constantly seeking lower cost methods which preserve as many of the quality-enhancing characteristics of personal visits as possible while controlling for interviewer bias. It has been little more than 20 years since the first attempts at CATI were made. CATI was followed by CAPI in the mid-1970s, then TDE in 1986, VR in 1987, CSAQ in 1988, FAX in 1991, and EDI in 1993. Internet-based collection began in the CES in 1996. Where will the next methodology come from?

A glimpse into the future may be available from the integration of networks that is already underway. Through the Internet, we will be able to send all survey messages, whether questionnaires, advance notices, nonresponse prompts, or edit reconciliation requests electronically. Use of E-mail will affect how both household and establishment surveys are conducted. Respondents could see the questions and answer them by keyboard, voice entry, or extracting data items from their electronic files in a "cut and paste" fashion.

Instantaneous transmission of data will allow on-line editing and simultaneous connection to interviewers to reconcile edit failures too complex for self-correction. By adding a very small video camera, the interviewer and respondent can see each other as in video-conferencing, providing the benefits of personal interviews with the quality control features and cost of CATI. By carefully defining the needs of selected respondent groups, whether for prompting, edit reconciliation, response analysis or re-education of replacement respondents, varying combinations of human and machine resources can be focused for maximum quality at the lowest cost. Such integrated systems would also allow for rapid collection of quick-response survey supplements addressing important economic issues. In addition to creating new methods, the next decade will see a continuing evolution of existing techniques, essentially re-engineering our current set of CASIC methods and adding new features and methods.

At this moment, in the middle of the year 2000, the worldwide web is the most salient frontier of information technology, including collection, transmission, and dissemination of survey data. With the addition of scripting languages such as VBScript and JavaScript, the implementation of cascading style sheets, and powerful encryption in standard web browsers, HTML has matured into a dynamic, interactive programming language with a great amount of control over screen display, edits, and question order. Users can connect to a survey site and enter their data into forms online, or web forms can be administered on a standalone computer at the interview site.

While the web is the present frontier, digital technology is exploding in many other directions, including increases in information processing capability on a single machine, and advances in connectivity between machines. As microchip technology has evolved since the invention of the transistor in 1947, we have seen exponential increases in processing speed and decreases in size. Also, with advances in superconducting, quantum computing, and other aspects of nanotechnology, upper bounds on speed and lower bounds on size cease to be limiting factors.

Connections between computers are becoming faster and more reliable and flexible as well; further, as more appliances and devices are built on digital foundations the term "computer" has come to apply to many household and business objects. Fax, telephone, wireless, email, and Internet capabilities are melding together to allow seamless communication across media. Special interface methods such as bar-code scanners, infrared ports, wearable computers, voice-recognition, active and passive styluses, virtual and augmented reality software and hardware, and other techniques suggest that the concept of the computer as an isolated machine is fading, and that computation will eventually be (sometimes literally) woven into the fabric of daily life. The implications for survey data collection are presently unimaginable; as a survey is simply a system for moving and manipulating information, all these new approaches have the potential for changing survey methods radically.

11. HOW DO CASIC INNOVATIONS OCCUR? ENTREPRENUERS AND SPONSORS

Each of these CASIC implementations represents a revolution, however small, in an existing series of production processes. It is interesting and useful to how these revolutions, or re-engineering, or innovations occurred and where they occurred organizationally. Where do CASIC initiatives come from? As mentioned above, virtually all such initiatives originate in the statistical methods divisions, not in systems, field offices or subject matter areas. This should not be surprising as the role of methodologists is to assess sources and magnitudes of errors. Methodologists also actively share these techniques through the annual ASA meetings and biennial research conferences.

The usual approach is the CASIC sponsor. Under the direction of a senior methodologist, a concerted effort to strongly encourage the research and development of techniques is seen. In the case of the CES, the CASIC entrepreneur was a senior statistical methodologist who became the program manager. In so doing, he sponsored the establishment of a separate branch-sized unit some what isolated from normal production, but with a clear mandate to develop and implement methods or methods needed to accelerate the timeliness of data collection, a very tightly measurable criteria. This branch, under a prodding and nurturing sponsorship, developed and implemented CATI, CAPI, TDE, VR, EDI and WWW collection. Can these approaches to innovation be replicated? What other approaches have been tried successfully to forward innovation and re-engineering of collection or other survey processes? Also, how does research get from inspiration to actual implementation?

In the CES case, each of the following features were present: a receptive research environment, a clear mandate to actually implement research results, sufficient funding, and a prodding and supportive, practical sponsor. This, "sponsor-entrepreneur" combination provided the impetus and momentum to overcome the barriers to change listed above.

Can an environment conducive of and supportive of ongoing innovation and re-engineering be established? The successes outlined here and in other organizations should be reviewed to identify the factors leading to improvements. It is clear that major improvements can be obtained in the right circumstances.

Figure A. ***Evolution of Uses of CASIC Methods at BLS***

Program	88	89	90	91	92	93	94	95	1994 units under CASIC	Est. 1999 units CASIC	Est. units in CASIC 2000	Est. units CASIC 2005	Considerating WWW
CES	CATI TDE					EDI; CATI NRP; FAX NRP	Fax for edits; Internet		25 %	100 %	85%	90%	In Collection
CPS				CATI									
					CAPI				100 %	100 %	100 %	100%	
MWR									- -	33 % (EDI)	5% of multis	20% of multis	√
SIC Refiling									- -	20 % (TDE)	0%	40% TDE/www	√
OES									- -	20 % (EDI)	3%	7%(diskette)	√
JOLTS											Starts April '00	100%	√
NLS		CAPI							100 %	100 %	100%	100%	
CPI	C&S								- -	- -	0	100%	
	Housing				Pen-Based				- -	100 %	100%	100% Pen-based	
PPI						CAPI			- -	100% multiple modes considered			√
TPOPS					CATI				- -	100 %			
CE									- -	- -			
IPP									- -	- -			
ECI									- -	25 % (EDI)	- -	50%	
EBS									- -	25 % (EDI)	- -	50%	
NCS									- -	25 % (EDI)	5%	50%	√
OSH									- -	- -	- -	15%	√
CFOI									- -	- -	- -	- -	
HAW		CATI for NRP							- -	- -		- -	

NEW METHODS FOR STATISTICAL PROCESSING IN A NEW ORGANIZATION ENVIRONMENT

W. J. Keller and A. J. Willeboordse, Statistics Netherlands

W.J. Keller, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands

wklr@cbs.nl

ABSTRACT

New IT-developments urge national statistical institutes to reconsider both their methods and organization with respect to data collection, editing, estimation, tabulation and dissemination. This paper states how Statistics Netherlands is preparing for the new era in making official statistics. It sketches Statistics Netherlands' overall view on IT-supported statistical processing, and how this view is currently being materialized in a new organization structure of the Bureau. Also, it elaborates on new methodology with respect to some crucial features of the modern survey climate: EDI-based *data collection*, followed by *estimation* of statistical outcomes on the basis of an often abstruse mixture of primary and secondary sources.

Key Words: EDI; Statistical processing, Process-integration, Organization structure, Estimation, Database, Repeated reweighting.

1. INTRODUCTION

The emergence of new information technologies urges statistical offices to critically review the methodology and organization of their statistical processes. Statistics Netherlands recently undertook such a review and concluded that a fundamental reorganization would be necessary, in order to take maximum benefit of the rewards of the IT-era. These benefits are manifold: respondent burden decreases, efficiency of statistical processing increases, and so will accessibility and coherence of statistical products. The core of the operation lies in the transformation from a subject matter oriented to a process-oriented organization. The new organization is due to be introduced in October 2000.

This paper comprises three main parts.

- First, an overall analysis of a technology based statistical process is presented. The process is described as a cycle, consisting of a number of phases that are delimitated by databases.
- Secondly, this view is translated in the process-based organization structure as is currently being worked out. Similarity of processes is the leading criterion for the new structure.
- Thirdly, an overview is given of new methods applying for the subsequent stages in the statistical process. The main focus is on EDI-supported primary and secondary data collection, as well as on one of the most intriguing challenges survey methodologists will encounter in the near future: how to produce an overall package of reliable and numerically consistent statistical information on the basis of a fragmented and incoherent bunch of data, supplied by a variety of primary and secondary sources.

2. AN IT-BASED CYCLE OF THE STATISTICAL PROCESS

2.1. The process as a cycle

The standard statistical process can be conceived as a cycle that is essentially similar to the production cycle of a composite and rather complex industrial product. The following stages and steps apply:

A. design stage

1. Exploration of user needs.
2. Exploration of production possibilities and cost.

After balancing 1 and 2:

3. Product design, resulting in a blueprint of the final product: a set of - empty - tabulations.
4. Specification of the production process:
   - *input*: choice of primary and secondary sources; form and sample design; administrative editing rules;
   - *throughput*: statistical editing rules; choice of methods for imputation; translation from input to output concepts; estimation; integration.

B. Implementation Stage

1. Input:
   - data collection / use of administrative registrations;
   - administrative/straightforward editing.
2. Throughput:
   - statistical/advanced editing; imputation; translation; micro-integration;
   - estimation;
   - meso/macro-integration.
3. Output:
   - tabulation;
   - publication;
   - dissemination.

2.2. The role of IT-tools

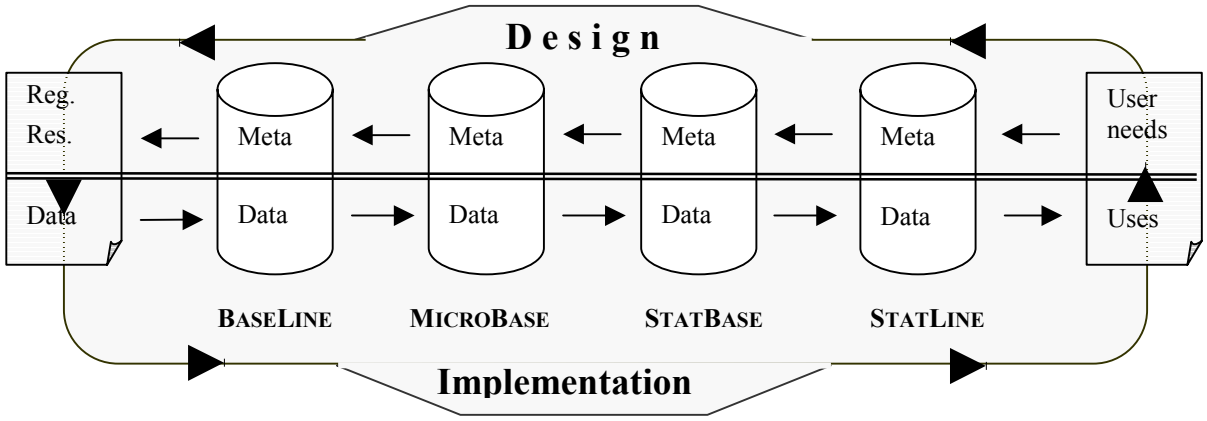
The statistical process cycle as described above rests on four pillars, notably *databases*, each of them representing a particular moment and delimitating a particular step in the process:

1. the input microdatabase BASELINE, holding all data as supplied by primary and secondary datasources. BASELINE eventually shows the data after administrative editing and mapping on statistical units;
2. the output microdatabase MICROBASE. It contains the data as they result from imputation, translation and micro-integration;
3. the output aggregate database STATBASE. It shows the results after estimation for (sub-) populations of statistical units. It can be claimed taht STATBASE holds *all publishable* data as produced by Statistics Netherlands;
4. the publication data warehouse STATLINE. It can be seen as *a set of views* on STATBASE, presenting the total output of the Bureau as a structured set of multi-dimensional tables, each representing an area of interest of the society. STATLINE is disseminated both on CD-ROM and on the Internet.

Above, the four databases were ranked according to their role in the *implementation* stage of the process. Notice that in the preceding *design* stage the order is the other way around: agreed user needs are first represented as - empty - STATLINE datacubes and specified in STATBASE in terms of metadata on statistical concepts and classifications. Subsequently, corresponding metadata are stored in MICROBASE, and finally input concepts are derived and stored in BASELINE.

The figure shows the joint design and implementation stage, together making a complete cycle, in which the metastream goes from right to left and, subsequently, the datastream from left to right:

Figure 1. The statistical process as a cycle



3. THE NEW ORGANIZATION OF THE STATISTICAL PROCESS

3.1. The main structure

The new structure is conceptually founded on the above outlined cycle analysis, in the understanding that *similarity of processes* is the predominant criterion for delineation of organizational units.
At the top level, four divisions occur, the first three of which are directly involved in the process:

1. Business Statistics (BES[1]);
2. Social and Spatial Statistics (SRS);
3. Macro-economic Statistics and Publication (MSP);
4. Technology and Facilities (TNF).

In terms of the cycle figure, one could roughly say that the first two Divisions take care of the part that is left of the database STATBASE, while the third deals with the part at the right side of STATBASE. The fourth one provides methodological, technological and domestic support to the others. The ANNEX depicts the new overall organization chart.

In this paper we confine ourselves to the organization units and aspects that are directly involved in statistical processing, i.e. the former three divisions and their departments. For a more comprehensive outline of Statistics Netherlands new organization structure see Willeboordse (2000).

3.2. Divisions Business Statistics (BES) and Social and Spatial Statistics (SRS)

The Division BES compiles statistical information on the activities of (groupings of) businesses and institutions, as well as on business-related aspects like technology, energy and environment. The Division SRS accounts for the demography, the activities and the living conditions of persons and households, and for spatial statistics. As their internal structure is essentially similar, the two divisions will be discussed simultaneously here.
In terms of the cycle outlined before, the goals of these sister-divisions can be phrased as follows:

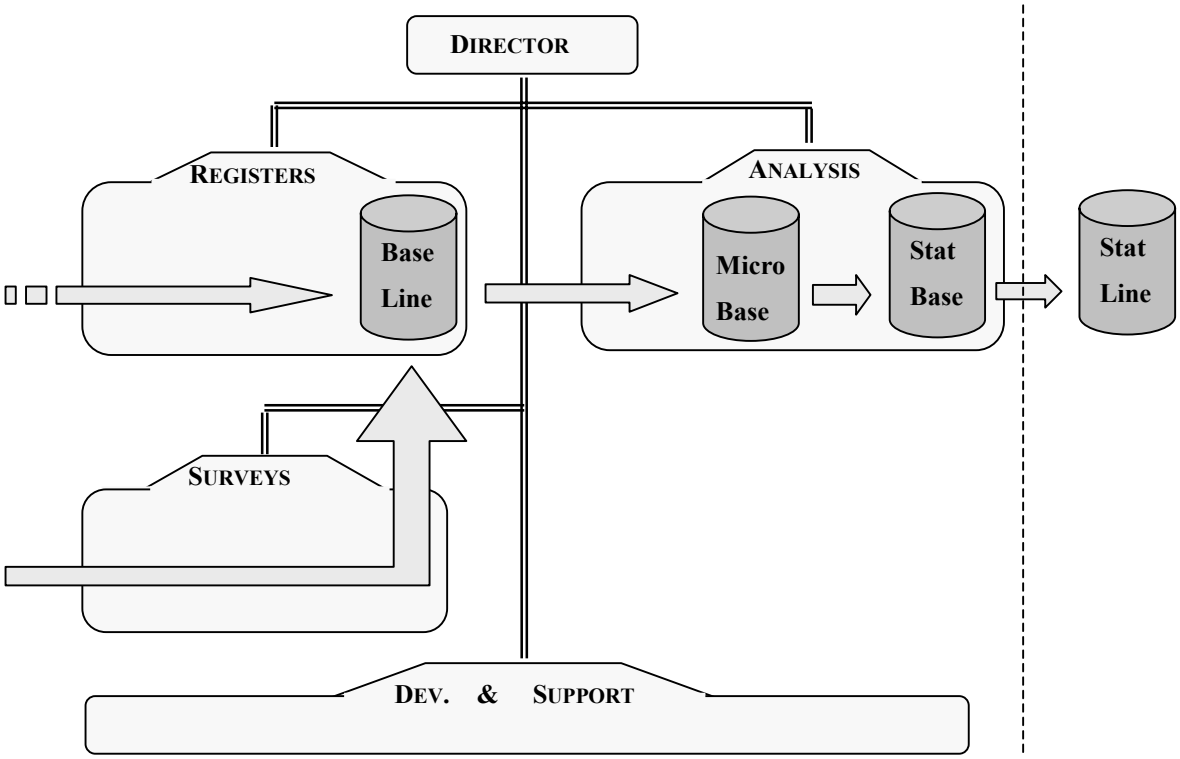
1. fill STATBASE with publishable data that comply with concepts and classifications as developed under the supervision of MSP. "Publishable" means: comprehensive, reliable, consistent and non-confidential;
2. contribute to the compiling of publications, to be issued by MSP;
3. provide MSP with inputs for macro-integration.

[1] Acronyms relate to the Dutch names.

Within the Divisions, the predominant distinction goes between data *collection* and data *analysis*, with the input database BASELINE as the demarcation point. The former demands skills that are primarily oriented to the world of the data sources, whereas the latter is more subject-matter oriented. With respect to *data collection*, a further distinction applies according to type of source, i.e. existing registers versus own surveys.

These considerations result in the following basic structure for each of the two divisions (arrows depict data streams):

Figure 2. Data streams in the Divisions for Business and Social Statistics



Department Registers

Both divisions have a REGISTERS unit, with the following tasks:

- *map* the data derived from a variety of registers on the relevant statistical units and *enter* units and data in BASELINE;
- *check* the data for obvious administrative errors;
- *Supply* SURVEYS with *sampling frames* for data collection.

Department Surveys

There is one such department in SRS, and there are two in BES. Main tasks are:

- *development* and *execution* of surveys, according to designs as provided by ANALYSIS;
- *editing*. The scope of the edits may vary with the type of statistics at hand. For specific subjects (e.g. environment) the editing is limited to obvious errors, while for general subjects (e.g. turnover) the editing includes relational checks;
- *supply* the data to REGISTER staff, which maps them on the statistical units in BASELINE.

To satisfy the one-counter principle, the departments are structured such that respondents are contacted by one sub-department only, whichever the type of survey. For the 250 largest businesses, this includes profiling of statistical units and attributes.

BASELINE

is the final product of REGISTERS and SURVEYS. It is the one and only source for the further processing of *all* statistical information by the two Divisions. It contains data on input-concepts, that are edited to a certain extent and mapped on statistical units, but not checked on consistency between different administrative sources or own surveys.

Department Statistical Analysis

Each of the two divisions has two ANALYSIS departments. They are the linking pins with the MSP division, both in the design and implementation stage of the process:

- in the *design* stage, they operationalize the working program as established by MSP and as embodied in the meta-component of STATBASE;
- in the *implementation* stage, they supply MSP with:
  - inputs for publications, by filling the data-component of STATBASE, and by contributing to publication texts;
  - inputs for macro-integration.

ANALYSIS bridges the gap between BASELINE and STATBASE. This comes down to transformation of input concepts for an incomplete and inconsistent set of "virtual respondents" to comprehensive, consistent, reliable and non-confidential, thus *publishable* data according to purported output concepts. This is done in a number of steps, each requiring specific statistical techniques: editing, imputation, translation, micro-integration and finally estimation.
The results after micro integration, so before estimation, appear in MICROBASE[2].

MICROBASE

records data on output concepts, at the level of individual statistical units. The data are obtained by editing and imputation of data provided by BASELINE. MICROBASE is the one and only basis for the estimation of aggregates to be entered in

STATBASE

being the final product of the Divisions BES and SRS. It contains publishable data on statistical concepts and classifications that satisfy the statistical program as developed by MSP.

2.1. Division Macro-economic Statistics and Publication (MSP)

Generally spoken, MSP accounts for stages in the statistical process that surpass the scope of Divisions SRS and BES res. Its role in the statistical process comprises three aspects, which are briefly discussed.

Macro integration

The Division produces a variety of integrated statistics, like National and Labour Accounts.

Programming: starting the cycle

The Division prepares Statistics Netherlands' *statistical program* and - after approval by the Central Commission of Statistics - materializes this program in terms of output metadata to be entered in STATBASE. As such, STATBASE entails the specification of the purported output, to be produced by the BES and SRS Divisions.

[2] Actually, there are several MicroBases: at least one for establishments/enterprises and one for persons/households, and later on probably more, such as for dwellings.

The concentration of all Statistics Netherlands' output metadata in one central database provides promising opportunities for the long cherished goal of *standardization* and *harmonization* of concepts and classifications.

Publications: closing the cycle

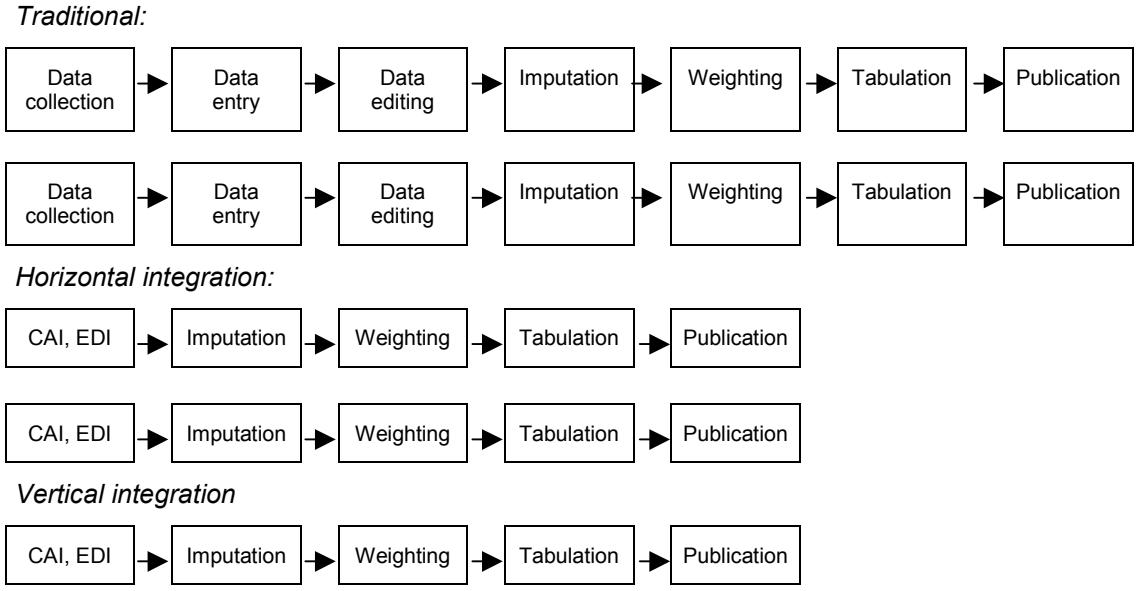
The Division is responsible for publication and dissemination of *all* statistical information, wherever produced in Statistics Netherlands. The output datawarehouse STATLINE is the spider in the web around which publication activities are organized. STATLINE consists of a number of datacubes, each describing a particular theme or part of a theme, and together covering the whole statistical program. All STATLINE data are derived from STATBASE; thematic cubes are *structured views* on STATBASE indeed. Themes are "owned" by *editors*, who are responsible for cube *design*, as well as for the *device* of all other publications in their area, including their share of the Internet website. The actual *preparation* of cubes and publications is the common responsibility of editors and data suppliers from the ANALYSIS departments in the BES and SRS divisions.

4. NEW TECHNOLOGIES AND METHODS FOR THE STATISTICAL PROCESS

This chapter goes in more detail with respect to the way advanced technologies and methods apply in the new organization environment as outlined above. The focus is on the use of EDI in *data collection* and on the exploration of new methodology regarding *estimation* on the basis of a variety of primary and secondary data sources.

4.1. New technologies for data collection

The traditional approach towards making statistics is based on data collection by means of (sample) *surveys* using *paper* questionnaires. Developments in IT have triggered a change from *paper data streams* to *bit data streams*. The paper form is gradually being replaced by an electronic one, or data are collected electronically without questionnaires and interviewers (EDI). This new way of data collection has a fundamental impact on the way statistical agencies operate. This impact reveals itself in two directions, as the figure shows. First, there is *task or horizontal integration* and secondly *survey* or *vertical integration:*



The new organization accommodates for vertical integration in both household and business surveys. In the process towards integration, we give priority to integrate data collection for business surveys that use EDI. This approach focuses on *data sources* (i.e. book keeping systems) instead of *surveys*: there is *one* (electronic) questionnaire for *one* book keeping system. Data collected from these systems may serve several statistics. Vertical integration means an important step towards the one-counter concept.

In persons/households surveys the share of CAI and EDI has evolved from 24% in 1987 to over 90% in 1998. In 2007 this percentage will approach 100, while two third of the CAI and EDI will refer to *secondary* sources, i.e.

existing administrative registers. In business statistics, the paper share is considerably higher: two third in 1998 and an estimated 25% in 2007. Here, we expect that the secondary sources will account for less than half of the EDI-share.

The introduction of EDI obviously triggered the currently undertaken *integration* of the organizational units involved in data collection. Actual integration of the major part of data collection for business surveys will, however, take a number of years. The speed of this development heavily depends on the pace with which the implementation of primary and secondary EDI evolves.

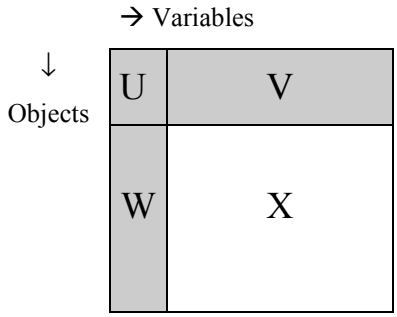
EDI has a substantial impact on the *methodology* of data processing activities as well. Some kind of *administrative editing* is needed at the input stage. This type of editing relates to detecting and correcting obvious *conceptual* and *processing* errors. It is characterized by batch-oriented error detection procedures. Detected errors are analyzed and corrected on a form-by-form basis. Correction may involve contact with the information supplier.

Finally, it should be stressed that EDI has beneficial effects on the *quality* of statistics as well, in particular with respect to consistency of concepts and data. Concentration of data collection leads to *standardization* of questionnaire concepts, and consequently creates better conditions for standardization of statistical concepts. Also, it triggers *consistent* reporting by respondents, which in turn creates better chances for consistent and relatable statistics. This is especially true for large and complex businesses.

4.2. Estimation: from MICROBASE to STATBASE

As stated before, more and more electronic data will become available as secondary sources for statistics. Important sources will be administrative (public) records in registers like those of the tax authorities and the social security administration, but also records from private sources. It is to be expected that in the future those registers will become the main data sources for statistical institutes. Combining the *abundant* amount of (cheap) register data with the *sparse* amount of (expensive) survey data in such a way that the results meet high quality standards will become a major challenge for survey methodology.

Figure 3. Survey and register data in MICROBASE



The figure depicts the differences between register and survey data, as they are stored in MICROBASE. The columns denote the variables to be observed and the rows list the objects (persons, businesses, etc.) for which the variables are measured. Most surveys are carried out by drawing samples that tend to cover *many* variables and a *limited* number of objects. Datasets U and V correspond to two sets of variables as observed in the sample, where U refers to variables that also appear in registers. Both sets of variables are defined by the statistical institute. Many variables are observed for few objects. For administrative registers, the opposite tends to be true: they cover *many* objects (often the whole target population) and a *limited* number of variables. In figure 3 the register is denoted by the datasets U and W. Neither definition nor observation is under control of the statistical institute.

Combining survey data with register data means matching the survey records to the corresponding register records. Matches must be established on the basis of the common part of the records, i.e. the measurements on the variables that are observed in both the register and the survey. Due to definition differences of these variables this may prove to be not an easy task.

After linking of registers, statistics can be compiled on the basis of the complete dataset. For statistics relating to variables in the set U + W, data are available on all population elements. So, there is no sampling error. The situation is more complex with respect to the variables in the set V. Here, only partial information is available. It is the challenge of survey methodology to predict the missing information in X, using a model for the relationship between the variables in U and V. In principle, this prediction can take two forms:

- *Mass imputation*. Based on models, synthetic values are computed for the missing data in X and actually entered in the data file;
- *Weighting*. Based on models, weights are computed and assigned to the complete records in U + V. In the computation of statistics only the weighted records in U + V are used.

The question which of the two methods applies best, places the survey methodologist in a true dilemma:

- Estimation methods based on mass imputation generally result in *consistent* estimates of related statistics. However, only a *limited* set of such estimates will meet *reliability* standards;
- Conversely, estimation methods that are based on weighting, will generally result in *reliable* estimates for *all* statistics. However, only a limited set of these statistics will show *consistent* data.

The strengths and weaknesses of both approaches are discussed in Kroese, Renssen and Trijssenaar (2000). To escape from the dilemma, they suggest a new approach. By *repeated reweighting*, they cope with the consistency problems of traditional weighting, without losing too much with respect to statistical reliability (see also Kooiman, Kroese and Renssen (2000). We will briefly explain their approach.

Two stages apply in the weighting process:

1. First, for each statistical figure to be estimated, the proper datasets as shown in the figure above are selected from the MICROBASE. If the statistical figures only concern variables that appear in the register, then U + W is chosen. Such figures can be estimated by straightforward counting up. Numerical consistency is guaranteed, since all estimates are obtained from the same list of records. Otherwise, if the statistical figures concern sampling variables or a combination of both sampling and register variables, then U + V is chosen. These figures are estimated by weighted counting.

   During this stage, consistency requirements are not yet predominant: a weighting model is chosen primarily to reflect the sampling design, a correction for possible non-response bias, and variance reduction. Still, the set of estimates based on weighted counting by itself is numerically consistent, since they represent the same underlying weighted microdataset. Furthermore, some consistency with register counts can be obtained, since already at this stage the weighting model is likely to involve calibration towards some (probably not all) register counts. Due to a fundamental lack of degrees of freedom it is generally impossible to use a weighting model that is sufficiently rich to guarantee numerical consistency with all register counts.

2. In the second stage, a so-called *minimal reweighting* procedure is applied for each statistical figure that has been inconsistently estimated in the first stage. This procedure is sequential and requires a careful examination of all relationships the variables involved in the actual table have with all tables that have been published before.

The example above illustrates the case of combining *one* register with *one* survey. In practice, focus should be on the more general case of combining many surveys with several registers, all containing measurements on the same objects. The method of repeated reweighting applies for these complex cases as well.

In the future, surveys should be designed such that the result of combining survey data with secondary source data is optimal. Again, it should be noted that developments towards integration have positive effects on the quality of the statistics. Needless to say that the combined treatment of various data sources, both surveys and registers, enhances relatability of statistical results. Matching operations reveal inconsistencies between different sources and provide the opportunity to solve them in an early stage of the process, in stead of having to wait until the integration in National Accounts or other integration frames.

The estimates resulting from the weighting operations as described above, are stored in the aggregate database STATBASE. The repeated reweighting method guarantees that the data are reliable and consistent. Also, methods must be applied to ensure that they meet requirements with respect to confidentiality. In the end, it can be said that STATBASE holds all *publishable* data Statistics Netherlands produces. As such, it is the final product of the Divisions for Business Statistics and Social & Spatial Statistics.

4.3. Tabulation: from STATBASE to STATLINE

STATBASE is the one and only source for the output datawarehouse STATLINE, which contains a number of multi-dimensional *datacubes*, each covering a theme or part of a theme and together providing a comprehensive and coherent picture of the society. As themes may overlap, the same data may appear in several cubes under different themes. STATLINE can be characterized as a standard view on STATBASE.

Designing a datacube such that it both covers a whole area of interest *and* shows a minimum number of blank cells *and* is easily accessible by users, requires new and specific skills of statisticians, which we refer to as the *art of cubism* (Altena and Willeboordse, 1997). In the new organization, this is the task of the *editors* ("theme-builders"), situated in the Department Publication of the SMP Division.

Datacubes may very well be huge-sized, both with respect to length, width and depth. Unlike paper tables, this does not lead to inconvenient presentations to the user. For, he is not confronted with a (huge) table, but with a *menu*, from which he can select the ranges of the table he is interested in. The challenge for the "artists" comes down to designing and structuring the menu, in such a way that the user can easily find and select what he is looking for, without having to shop around in different cubes in order to satisfy his data needs.

REFERENCES

Altena, J.W. and A.J. Willeboordse (1997): Matrixkunde of "The Art of Cubism" (Dutch only). Statistics Netherlands, Voorburg.

Bethlehem, J.G. and F. Van der Pol (1998): The Future of Data Editing. In: M.P. Couper et al.: Computer Assisted Survey Information Collection. Wiley, New York, pp. 201-222.

Bethlehem, J.G., J.P. Kent, A.J. Willeboordse and W. Ypma (1999): On the use of metadata in statistical processing. Third Conference on Output Databases, Canberra, March 1999.

De Bolster, G.W. and K.J. Metz (1997): The TELER-EDISENT Project, Netherlands Official Statistics, Autumn 1997, Statistics Netherlands, Voorburg.

Keller, W.J. and J.G. Bethlehem (1998): The impact of EDI on Statistical data processing. Meeting on the Management and Information Technology, Geneva, February 1999.

Keller, W.J. and A.J. Willeboordse and W.F. Ypma (1999): Statistical Processing in the New Millennium. Proceedings of Statistics Canada Symposium 99, Combining Data from Different Sources, Ottawa, May 1999.

Kooiman, P., A.H. Kroese and R.H. Renssen (2000): Official Statistics: an estimation strategy for the IT-era. XIV Conference of the International Association for Statistical Computation, Utrecht, August 2000.

Willeboordse, A.J. (2000): Towards a New Statistics Netherlands. Netherlands Official Statistics, Spring 2000, Statistics Netherlands, Voorburg.

Kroese, A.H., R.H. Renssen and M. Trijssenaar (2000): Weighting or Imputation? Netherlands Official Statistics, Spring 2000, Statistics Netherlands, Voorburg.

ANNEX Organization Chart of Statistics Netherlands (October 2000)

Total staff in 2003: 1980

numbers are provisional