Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 3–21

Information Technology and Survey Research: Where Do We Go From Here?

J. Merrill Shanks[1]

Abstract: This article reviews the areas in which information technology has had an impact on the cost, quality, or complexity of survey research, and discusses alternative strategies for integrating the computer-based activities which take place in different stages of the survey research process. Special attention is given to the continuing revolution in survey data collection, for it is the only area of applied computing that is unique to the survey field and is central to the eventual integration of data collection, analysis, and management.

Key words: Computer-assisted surveys; data collection; data management; data analysis; computer-assisted telephone interviewing (CATI); generalization of technical procedures; alternative computing environments; survey integration.

1. Introduction

Forty years have passed since the U.S. Bureau of the Census acquired the first UNIVAC to process data based on structured questionnaires. Since that time, the list of activities which depend on computers for the collection, processing, or management of survey-related information has grown steadily, to the point where nearly all aspects of the survey process are now at least partially dependent on computer-related technology. This article begins with an overview of survey activities where computers have already had an effect, with an emphasis on the current revolution in computer-assisted data collection. Most of the article, however, concerns alternative strategies for future research and development and the integration of separate techniques for data collection, analysis, and management.

Survey organizations must make increasingly complex (and expensive) decisions if they are to take advantage of continuing advances in computing and information technology. Survey projects that either develop or use computer-based techniques have a tendency to concentrate on short-term improvements or immediate objectives. The purpose of this article, however, is to re-emphasize long-term objectives in computer-assisted surveys and discuss alternative strategies for future development in that field. In cooperation with other survey organizations, the Computer-assisted Survey Methods

[1]Director, Computer-assisted Survey Methods Program, University of California, 2538 Channing Way, Berkeley, CA 94720, U.S.A.

Program (CSM) at the University of California is active in several aspects of the relationship between "computing" and "surveys." As a consequence, the observations which follow represent both a general commentary on our field's progress to date and specific recommendations for a research and development agenda that is shared by many other organizations.[2]

1.1. Computer-assisted surveys: Rationale and impact

For several years, the author has been a frequent visitor in other survey organizations, in order to discuss computer-based systems for various aspects of survey research. Most of these visits have been initiated by organizations that are interested in computer-assisted telephone interviewing (CATI), but the resulting discussions almost invariably cover a variety of other computer-related activities. At first, it seemed reasonable to assume that all participants in such discussions shared the same basic objectives in adopting new computer-assisted techniques. It quickly became clear, however, that the motivations and expectations which precede the adoption of a computer-based system can vary substantially within the survey organization. Our field is no different from many others in this respect, for potential users may be attracted to any computer-based system for one or more of the following (partially contradictory) reasons:

- the output from the process may be better, in that the resulting information will be more accurate or of higher quality;
- the entire process may be less expensive, even though computing equipment represents a cost that had not previously been involved;
- the entire process may be faster – whether or not it is less expensive, for time and money represent two different kinds of "resources;" or
- the process may be more powerful, for the task or the design may be so complex that it could not be done accurately without a computer-based system.

These kinds of expectations represent the intended consequences for survey researchers who consider the adoption of new computer-based techniques. Such adoptions, however, often lead to other (unintended) consequences, for significant changes can take place in the division of labor and structure of an organization after new technology is introduced.

To be sure, relatively few survey organizations adopt computer-based systems in order to change their bureaucratic structure or division of labor. As in other fields, however, new technology is often introduced for reasons that are partially in conflict, and groups with different objectives may compete for control of the new procedures. Frequently, one group or fraction seeks faster results or lower costs, while another emphasizes the possibility of greater quality or complexity – either of which may require additional funding or time.

In addition to conflicts between objectives, organizations that adopt computer-based systems frequently experience shifts in their inter-

[2]The opinions expressed in this article are entirely the author's, but they are based on cooperative projects with specialists in survey research from several institutions. In particular, the CSM Program has been working with the National Agricultural Statistical Service (NASS) since 1981 to develop, test, and utilize general-purpose systems for the collection and processing of survey data. The CSM software used in that project has been developed in an unusually collaborative fashion, involving staff members from the University of California, NASS, the Bureau of Labor Statistics (BLS), and several dozen other organizations that participate in the Association for Computer-assisted Surveys. The author is indebted to several colleagues for their ideas and criticism, but the views expressed in this article are not necessarily shared by other individuals in any of the organizations involved.

nal structure to fit the procedures that are defined by the new technology. Such changes were first evident during the 1960s and 1970s when survey researchers began to write their own statistical programs instead of relying on their centralized data processing staff. In the current period, the impact of technology on survey organizations has been particularly visible in the trend toward computer-assisted data collection.

Each organization is somewhat different in this respect, for survey projects vary enormously in their complexity and division of labor. Despite such differences, however, computer-based systems for data collection usually lead to some form of organizational change, for they combine survey activities that were previously carried out in separate offices or groups – i.e., in sub-organizations that specialize in instrument design, sampling, field, coding, or data processing. As additional survey activities are converted to rely on the same system or database, technical integration is required at the study or project level that simply did not exist before the shift to computer-assisted data collection. As emphasized below, this trend toward technical integration is likely to accelerate, as survey organizations improve linkages between separate systems for data collection, management, and analysis. Each step toward integration may suggest further changes in division of labor, as well as making some progress toward the objectives listed above – i.e., making surveys better, cheaper, faster, or more powerful.

The following section reviews the variety of information processing activities where computers have been used in survey research, with an emphasis on the changes that are still taking place in data collection – i.e., in the production, recording, and editing of survey information. The rest of the article then discusses alternative strategies for future development, with an emphasis on data management and the technical integration of the entire survey process.

2. Computer Applications and Survey Procedures

Many researchers can recall when survey data were stored on punched cards and processed by electromechanical devices instead of computers. It therefore seems natural to begin by identifying the range of traditional survey activities that involve some kind of "information processing" and the way in which computers were introduced in each of those areas. What follows is an oversimplification in some respects, but it suggests the variety of survey-related activities that may be affected by computing technology. (See Sonquist (1977) for a comprehensive review of computer applications in survey research.)

As shown in Fig. 1, most traditional survey activities have long been at least partially converted to computer-based procedures. The first area in which computers were widely used was the statistical summarization (or analysis) of coded data that had already been transferred from interview protocols to punched cards – which appears in Fig.1 as the penultimate activity in the traditional sequence from design through collection to post-survey data management. The first statistical programs were designed for a single type of computation (e.g., "means," "cross tabulation," or "correlations"), but software in this area quickly evolved into comprehensive packages, some of which now reflect over two decades of continuous development.

It is difficult to exaggerate the extent to which computers have improved the process of analyzing survey data. We now routinely expect results to be produced and displayed in a few seconds that once required an entire staff of technicians at desk calculators or unit record equipment. Survey researchers sometimes complain that our analytic strategies have not kept pace with improvements in the speed of computation. There can be no doubt, however, that survey analyses are now routinely completed

that are far more accurate, cheaper, faster, and more complex than was possible before the introduction of digital computers. Specific capabilities in this area will not be discussed in this article for the systems involved are frequently reviewed in statistical or computer-related publications. For example, the *Statistical Software Newsletter*, published in connection with the International Statistical Institute, is entirely devoted to reviews and critiques of statistical software. For more comprehensive reviews, see Francis (1981) and Raskins (1989).

The traditional division of labor in survey organizations involves separate staffs for sampling, field, coding, data processing, and analysis. Because of these structures, computer applications for the other activities described in Fig. 1 were usually developed independently of statistical software, and were often based on different computers and data processing conventions. The piecemeal or uncoordinated nature of these initial computer applications can also be partially attributed to the sequence in which new computing capabilities became available. As shown in Fig. 2, survey researchers have worked with a series of "new" infor-

Fig. 1. Information-processing Activities in Traditional[3] Survey Research

Original Activity	*Initial Computer Utilization*
Preparation of Interview Schedules and Specifications	Text Processing and Document Preparation, for both Questionnaires and Interviewer Instructions
Sample Selection and Administration	Data Management, for Sample Selection and Preparation of Pre-interview Data
Interviewer Supervision and Management	Keeping Track of Field Outcomes and Measurement of Interviewer Performance
Content Analysis for Text or Verbatim Responses	Conversion of Coded Data to Machine-Readable Form (Keypunching or Direct Data Entry)
Data Preparation: Checking and Documenting the Resulting Data	Detection and Correction of Inconsistencies and Preparation of Machine Readable Codebooks
Survey Analysis: Producing Tables and Charts	Statistical Computation, including Graphic Displays as well as Descriptive and Inferential Statistics
Post-Survey Data Management: Storage and Retrieval	Generation of Composite Data Files, and the Development of Data Archives

[3]For this discussion, the term "traditional" is intended to suggest personal (or face to face) interviews in which questions are read from a structured questionnaire and answers are recorded on the same printed form. Survey data are of course also collected without computers through telephone interviews and self-administered questionnaires. The above list is designed to suggest the range of "information processing" activities that were originally carried out without computers, as well as the ways in which computers were first used for those activities.

mation processing technologies, most of which simply converted a narrow set of activities to computer-based processing without affecting other areas or stages in the survey process. For example, programs for data management and statistical analysis were well established before packages became widely available for text processing or document preparation. For this reason, procedures for handling survey-related text – e.g, questionnaires, instructions, or reports – were typically developed on computers and by individuals who worked independently of those with responsibility for collecting or processing the resulting data. Similarly, procedures for documenting and archiving survey data were not effectively linked to major systems for statistical analysis, so that most survey analysts still cannot easily generate printed results that include the full text of the questions or procedures that were used to collect the data.

The data management tools required for large scale survey operations were also developed independently of statistical software. As a result, the character-oriented formats used for the collection and storage of survey data often conflicted with the requirements for numeric representation imposed by early statistical packages. Subsequent developments in both areas have converged in many respects, including compatible data formats as well as overlapping capabilities. Thus, several statistical packages can now handle more complex structures, and "database" systems have acquired some statistical capabilities. As discussed below, however, essential differences persist between modern systems for data base management and statistical analysis – and both differ significantly from software developed for questionnaire-based data collection and archiving – so that new strategies will be needed for linking or combining these separate technologies.

The most important changes now taking place in computer-assisted surveys stem from the revolution still taking place in data collection. During the 1970s, experimental programs were introduced which displayed questions and accepted responses at terminals operated by telephone interviewers. As described below, systems for Computer-assisted Telephone Interviewing (or CATI) quickly grew to handle several other types of "information" associated with telephone surveys, and the resulting systems have been generalized to the point where

Fig. 2. Information Technologies Used in Survey Research (in Approximate Chronological Sequence)

Initial Technology	*Extensions and/or Generalizations*
Unit Record Equipment: Devices for Processing Data on Punched Cards	Replacement of Punched Cards by Magnetic Tape and Disk Storage
Statistical Programs: Faster Computation Than Unit Record Equipment	Comprehensive Statistical Packages (e.g., BMDP, OSIRIS, SPSS, SAS)
Data Management: Utilities for Updating and Manipulating Files	General Purpose Systems for Relational Database Management
Text Processing: Using Computers to Process Text as Well as Numbers	Utilization of the Same Text in Questionnaires, Codebooks, and Reports
Data Capture: Separate Systems for Telephone Interviewing and Data Entry	General Systems for Data Collection Based on Structured Questionnaires

they can be used for other types of questionnaire-based data collection – and in a variety of computing environments. For an earlier collection of essays on computer-assisted data collection, see Freeman and Shanks (1983). See Nicholls and Groves (1986 a and b) for a more recent review of computer-assisted telephone interviewing, and see Shanks and Tortora (1985) for a discussion of the specific approach to CATI and its generalization being followed by members of the Association for Computer-Assisted Surveys.

2.1. *Stages in the development of computer-assisted surveys*

The resulting systems for data collection have also reached a stage in which survey researchers are considering the possibility – and the potential advantages – of integrating all of the information processing activities involved in the survey research process. For example, a computer-assisted telephone survey may rely on the same computing environment for document preparation (to create the interview schedule and interviewer instructions), data management (to handle survey information that is collected outside the interview context), and statistical computation (to describe progress or problems in sample completion) – as well as production interviewing.

The range of activities carried out in the same computing environment has encouraged speculation about a unified – or comprehensive – approach to the entire survey process, in which unnecessary duplication of effort might be eliminated without sacrificing any existing capabilities for data management, collection, or analysis. The ultimate objective for integration of this sort is a reformulation of the entire survey process, in which researchers will be able to concentrate on the content and quality of the resulting information – rather than on complications arising from the information processing environment.

Survey researchers have only recently started to work on this kind of technical integration. As in other fields, computer-based development projects in survey research can be assigned to one of three distinct stages, depending on whether they concentrate on:

- the development of initial programs for a specific applications and computing environment;
- the generalization of systems to related activities and alternative computing environments; or
- the integration of multiple systems for different activities, including linkages between systems based on different approaches.

At the present, however, most developmental projects in computer-assisted surveys can be classified in the first or second of these categories, for much remains to be done to improve and generalize systems for data collection, analysis, and management.

The following paragraphs concentrate on the improvement – and generalization – of systems for data collection, because those systems represent the only computer application that is unique (or indigenous) to survey research – and because data collection procedures will have a pervasive influence on the technical integration of the entire survey process.

2.2. *Computer-assisted data collection*

As suggested above, computer programs that were originally developed for telephone interviewing (CATI) have evolved into systems that handle many telephone survey activities in addition to administering the questionnaire. Because of that evolution, the same systems are also being generalized to handle a variety of other forms of data collection. The reasons for that generalization, and for the growing accep-

tance of CATI-type technology, can be seen in the comprehensive nature of the activities involved. Figure 3 reviews the ways in which telephone surveys may be affected by a CATI system. Figure 3 is similar to several published lists of CATI-related activities. See Shanks and Tortora (1985) for an earlier summary of this sort as well as references to other discussions.

Fig. 3. Telephone Survey Activities Affected by CATI

Preparation of the Interviewer's Instrument – drafting complete specifications for question content, question sequence or branching, and interviewer instructions, and entering those specifications into the computer;

Translation and Checking of the Interviewer's Instrument – transforming the computer-based instrument into a format which maximizes efficiency in interviewing, and checking all specifications for syntax errors;

Creation of Sample File(s) and Scheduling Instructions – creating a computer-based data set which contains a record for each case with telephone numbers and/or other identifying information, data from previous interviews, random numbers to control assignment to alternative question sequences, and information to be used in scheduling calls;

Study Management – producing periodic reports on study progress, interviewer performance, and sample completion, as well as assignment of calls to specific interviewers;

Production Interviewing – includes repeated dialing using assigned search patterns to establish contact with eligible respondents and the routing of problematic cases to supervisors for special handling, as well as actual interviewing;

Interviewer Supervision – resolving cases where interview attempts have been unsuccessful (through reassignment to language or refusal specialists, or to a final non-interview status), monitoring interviewer performance, and assisting interviewers on request;

Specification of Coding and Cleaning Procedures – preparing instructions to editors (or coders) and to the computer to control any checking, cleaning, or supplementary data entry which should take place after each interview is complete;

Conversion and Checking of Coder's Instrument – a process which may resemble translation of the interviewer's instrument (above) if the instructions for cleaning (or coding) are stored in the same format;

Production Coding and Cleaning – assigning coded values to unstructured text associated with open-ended questions or "other specify" responses, and resolving any inconsistencies between recorded responses and the logic of the (coding and cleaning) instrument;

Certification and Output for Completed Cases – final checking for errors in the data and transferring satisfactory cases to an output file for analysis;

Data Analysis and Documentation – using the information in the interview (or coding) instrument to produce explanatory text for statistical reports and final survey documentation.

Since the first CATI systems were introduced, lists of this sort have suggested that the new technology might incorporate (and thereby integrate) activities that were traditionally handled by separate groups or staffs. In particular, the computer-based "instrument" that controls a given CATI application may include instructions for activities that were traditionally carried out separately by specialists in: questionnaire design, sampling, interviewing, coding, supervision, data preparation, analysis, and archiving. By incorporating instructions for several of these (previously separate) activities into a single computer-based instrument, CATI projects can make several information processing activities do double or triple duty. The following examples illustrate this (now familiar) potential for consolidating previously distinct activities:

- the same computer-based files may be used to define the sample, control the sequence in which cases are assigned to interviewers, and provide documentation concerning the progress or history of data collection for each case;
- the test and logic of the interviewers' instrument may be converted into a parallel instrument for controlling all post-interview data entry and definition, as well as documenting the final dataset;
- answers or response patterns that are defined as illegal need not be corrected after initial data collection, for such errors are detected (and resolved) during the interview;
- the same instrument may be used to ensure that all appropriate questions are answered, even if the interviewer (or coder) has deviated from the prescribed question order by skipping ahead, moving backward, or changing an answer;
- no separate process is needed to convert interview responses to machine-readable form, since all data (including precoded responses and verbatim text) are "captured" during the interview through direct keyboard entry.

The potential advantages of this kind of consolidation are responsible for the rapid growth and dissemination of CATI systems. Continuing growth in both capabilities and usage has also led to the generalization of such systems, both to multiple computing environments and to other forms of data collection. In addition to this process of generalization, however, CATI systems are still being changed frequently, for much remains to be done before telephone surveys make efficient use of current technology for all the activities mentioned above.

For example, the CSM program is currently concentrating on several CATI-related enhancements to the Computer Assisted Survey Execution System (or CASES), including: computation and storage for multiple types of variables (including floating point), additional kinds of screens and forms-type processing, automatic scheduling of telephone interviews, transfer of cases between computers (for distributed data collection), more efficient data storage, and "help" facilities to make it easier to use all of the programs involved. These new capabilities are sometimes released individually, to meet the needs of specific projects or users, but they are usually combined into major versions or releases. As of this writing, CASES users are testing Version 3.3E, and plans exist for three more (major) versions before all of the currently scheduled enhancements are completed. Informal reports suggest that other data collection systems (besides CASES) are going through a similar process of revision or enhancement, so that many survey organizations experience frequent changes in their computer-assisted data collection procedures.

2.3. Generalization to alternative types of data collection

The above kinds of changes represent important enhancements for many CATI users, but they have had to compete for developmental resources with a quite different set of objectives based on the general nature of the activities involved. The breadth or diversity of any system's user community is an important determinant of the resources it can devote to development and maintenance. For that reason, and at the request of specific users, several CATI systems have been revised so that the same program can be used in an wider variety of contexts.

The first of these types of generalization (and revision) stems from the understandable desire of many survey organizations to use the same kind of procedures for projects which use different types (or modes) of data collection, and for single projects that must use more than one of those modes, including:

- Computer-assisted Personal Interviewing (CAPI),
- Self-Administered Questionnaires (SAQ), for Respondent-Entered Data, and
- Direct Data Entry (DDE), for Paper-and-Pencil Forms, as well as
- Computer-assisted Telephone Interviewing (CATI).

Two of these extensions (CAPI and SAQ) are currently limited because of their requirement that respondents (or subjects) be brought into direct contact with a computer.[4] With the continuing improvement in portable computers and communications, however, self-administered options may become much more important for several types of research. For example, computer-based questionnaires are already being administered on the telephone without an interviewer. In this approach, questions are presented through voice reproduction to respondents who call a designated number on a touch-tone phone. The respondent then answers the (voice reproduction-based) questions by entering numeric codes on the phone. This technique is called touch-tone data entry (or TDE) by researchers at the Bureau of Labor Statistics, where it is being used as an alternative form of data capture (to be combined with telephone interviews conducted in CASES) for the Current Employment Survey. (See Working, Tupek, and Clayton (1988).) Also, voice- and graphics-oriented options will soon be available for applications in which the respondent (or subject) can interact directly with a computer, instead of over the telephone. By simply "calling" other programs or devices, structured questionnaires may soon take on a very different character, as the concepts of "question" and "response" expand to include both images and sounds.

2.4. Generalization for distributed data collection

Most CATI systems were originally developed for a single (multi-user) computer, in which interviewers sat in front of terminals connected to the computer by direct lines or telephone.

[4]The generalization of CATI to Computer-assisted Personal Interviews (or CAPI) is still a moderately recent development, and several approaches are being explored to integrate the (computer-assisted) questionnaries in projects which call for both CATI and CAPI. Organizations working with CASES are developing similarly structured (but separate) instruments for each data collection method, based on the assumption that differences between modes (in instructions or logic) cannot be handled in a fixed or system-prescribed fashion. In contrast, the BLAISE system being developed by the Central Bureau of Statistics in the Netherlands can be used to produce a single instrument that is processed in a different way for CATI, CAPI, or direct data entry (DDE) based on paper forms. See Denteneer, Bethlehem, Hundepool, and Keller (1987).

For some time, the only exception to this rule was the Wisconsin system for micro computers, but PC-oriented survey organizations can now choose between many systems or approaches.[5] Increasingly, however, many survey projects require that data collection facilities be "distributed" across several computers, in one or more of the following ways:

- personal (or single user) computers are used for interviewing, but all of the data is maintained by a single file server over a local area network;
- a multi-user system serves as a satellite to a master (or host) computer (i.e., hierarchical relationships may exist between multi-user computers within a single facility); or
- computing facilities that are geographically (and organizationally) separate must be centrally coordinated for a specific project.

Within the CSM user community, each of these approaches to distributed data collection is already in use. The computer programs involved, however, need substantial changes to more effectively carry out (and check) the inter-system communications involved. The general problems of maintaining study-level integrity during data transfer between computers can be particularly severe when the two (linked) systems have different hardware and operating systems. An early requirement, therefore, for some applications has been that the programs involved function the same in several computing environments.[6]

The most important developmental tasks, however, have only begun, for data collection systems need more sophisticated protocols for transferring information from study-level databases between computers – regardless of the hardware and operating systems involved.[7]

3. Data Management and Survey Integration

Since the mid 1970s, survey organizations have concentrated most of their resources for computer-related development on general-purpose systems for collecting data based on structured questionnaires. The resulting programs are no longer new or experimental, for they are routinely used for "production" data collection in projects based on self-administered questionnaires as well as telephone and personal interviewing. As indicated in the previous section, however, much remains to be done in improving and generalizing those systems before they satisfy all of the objectives which have been identified by their user communities.

While those systems are still being improved, survey organizations have also become interested in the capabilities for handling large and complex data structures offered by systems that were developed for management – rather than the collection or analysis – of survey-type information. Survey activities which may call for a separate data management system include:

- questionnaires with more complex relationships than simple hierarchies (including "many-to-many" relationships like those between multiple patients and doctors);

[5] See Palit and Sharp (1983) for a statement of objectives for the Wisconsin system. Other data collection systems for the PC environment include those produced by Sawtooth, Inc. and Computers for Marketing, Inc., as well as CASES and BLAISE.

[6] For example, CASES programs have been converted for use in VMS (for VAX systems produced by the Digital Equipment Corporation) and a variety of UNIX systems in addition to personal computers that use PC- or MS-DOS. Work has also begun on an MVS version for IBM-compatible mainframe systems, and a version is planned for MacIntosh (Apple) computers.

[7] See Statistics Sweden (1989) for a description of their approach to distributed data collection which involves personal computers in interviewers' homes.

- interlocking surveys or "studies" in which more than one questionnaire may be administered in one interview with the same respondent;
- the management or allocation of data collection resources between multiple (simultaneous) surveys, and the measurement of staff performance across survey projects;
- the creation of large or complex datasets by combining information from multiple (survey and non-survey) sources; and
- the creation and maintenance of data archives, or comprehensive collections of datasets and documentation for a large number of surveys in a general area.

For these and other reasons, survey researchers are now exploring the potential benefits of "database" technology in managing complex data structures and integrating information from multiple sources – while continuing to rely on existing systems for data collection, analysis, and documentation.

3.1. Distinguishing data management from data collection and analysis

The number of different sources of survey-related "information," and the relationships between those sources, present a classic illustration of the circumstances in which an organization may benefit from using a relational database management system (or RDBMS). For example, a survey organization may already be maintaining separate computer-based files containing information about the following kinds of "entities" in addition to the data being collected or analyzed:

- past instruments (including both question wording and interviewer instructions);
- staff members (including employment history, hourly costs, and previous performance, as well as hours spent on each current project);
- sample elements (including information about unused cases as well as those assigned to current studies);
- multiple projects (including administrative information such as planned expenses vs. actual costs, as well as personnel plans and time schedules); and
- completed datasets (including documentation concerning data type and location of variables, time period, access permissions, and relevant publications).

A single data base application could include files for each of the above types of entities, so that users in one area (or project) would have access to information that was originally collected for other purposes. In applications of this sort, the data management system must support all of the linkages or relationships involved (e.g., between such entities as projects, instruments, staff members, sample elements, datasets, and variables), and it must permit users to define their own reports for retrieving and displaying information. The central concept in relational data base technology is the decomposition of any application into a series of simple or rectangular datasets (one for each type of entity), each of which is linked to other datasets through relationships between the entities involved, such as membership in the same family or data collection project.[8]

While survey researchers explore a variety of RDBMS applications, basic systems will continue to improve for data collection and analy-

[8] See Codd (1970) for an influential summary of design principles for relational database management systems. See Baker (1987) for a lucid account of the ways in which these concepts can be used to improve the management of traditional survey operations (based on household samples and face to face interviews).

sis, as well as database management. Most of the gains in survey-related computing will therefore continue to be extensions of single-purpose systems, i.e., as additional features in packages that were originally designed for data collection, analysis, *or* management – rather than entirely new systems which carry out all three kinds of activities.

To some extent, the boundaries between systems for data collection, management and analysis are becoming less distinct, for the major packages in each area have acquired capabilities in other areas without sacrificing the integrity of their original applications. Thus, statistical (or analysis) systems have acquired options for data entry and handling non-rectangular stuctures, and database management systems can be used for data entry and statistical calculations as well as displaying characteristics of individual cases. Similarly, as discussed above, systems for computer-assisted telephone interviewing (CATI) have been adapted for other (non-telephone) forms of data collection, and may include capabilities for statistical analysis and data management. This expansion of existing systems across the three basic "stages" (collection, management, and analysis) will continue for some time, but it is unlikely to produce a satisfactory computer-based integration of the entire survey research process. Experience to date suggests that the combined set of information-processing activities in projects based on structured questionnaires is extremely diverse. No single system (for data collection, analysis, *or* management) will soon reach the point where it provides all of the capabilities required.

As suggested in the introduction to this essay, each stage in the survey process is characterized by its own information processing requirements and complexities, many of which have been "handled" by simply ignoring information that is essential at other stages. For example, data collection procedures rest on complex instructions concerning the sequence in which steps are to be taken (or repeated) as well as voluminous instructions to staff members involved in the data collection process. Current practices in the management and documentation of survey data incorporate only a portion of those instructions, and almost all of that detail is discarded when creating analysis files for most statistical packages. Similarly, data management systems emphasize the relationships between entities and fields in different files, but such systems are usually intensive to the sequence in which data values should be entered or (re-)calculated – and they retain very little of the information about data content or the collection process that is typically included in survey data documentation (or codebooks).

In effect, each type of system has concentrated on a distinctive type or aspect of survey data processing, while disregarding information and logic which may be crucial at other stages of the research process. These differences, or simplifications, have made it easier to develop our existing systems for data collection, analysis, and management. The resulting differences, however, can make the databases produced by these separate technologies very difficult to integrate or combine.

3.2. Barriers to integration: Differences between stages and systems

Within any survey organization, one of the most frequent kinds of computer-related irritation arises from the utilization of several different systems for data collection, analysis, or management. As specific systems are generalized to the point where they can be used in different operating systems (and computing hardware), some difficulties in transferring data between systems may disappear – for the same computing environment can be used for different activities or stages. Barriers to data transfer or integration, however, can still arise when all activities (i.e., data collection, analy-

sis, and management) rely on the same computing environment, because of dissimilarities between the data structures used by packages that were originally developed for different purposes. In particular, systems for data collection, management, and analysis rest on quite different strategies for representing the instructions used to collect survey information and the content (or structure) of the resulting data.

This problem is compounded by the existence of alternative systems within each of the major stages in the survey research process. Survey organizations now rely on a substantial number of alternative systems for data collection and analysis, and several relational database management systems are now being evaluated for handling large or complex data structures. Any comprehensive list of systems now in use in the United States would include more than a dozen packages for CATI-type operations and many more than that for statistical analysis. In the foreseeable future, it is unlikely that any single system (for collection, analysis, *or* management) will grow to the point where it includes all of the capabilities needed in the other two areas, and multiple alternatives are certain to exist within each area for many years to come. Given that prospect, the inherent problems of integrating survey activities between the three major stages of the research process are compounded by the sheer number of system-to-system linkages involved.

To be sure, a single survey project may use only one system for data collection, one for statistical analysis, and a third for database management, so that it might need only two or three (bilateral) conversions of data from one system to another. Many organizations, however, use more than one package in some areas, and each system-to-system linkage (or conversion) can involve a substantial investment in software development. As a consequence, the developmental effort required for a single organization to transfer information between systems may be quite large – and must be repeated in organizations that use other combinations of packages – unless a more general solution to the problem can be found.

3.3. Alternative strategies for technical integration

Attempts to transfer computer-based survey information between systems encounter a variety of difficulties, based on differences in data structure between alternative systems in the same general category (or stage) as well as general differences between stages in the type of survey information involved. Most such problems can of course be "solved" through additional programming, but such projects can be very time-consuming and expensive. For this reason, researchers in several organizations have expressed interest in a general-purpose approach, so that solutions developed in one context can be used by other projects or organizations.

The barriers to linking or combining computer-based information maintained by different systems (for data collection, analysis, and management) represent what might be called the "last frontier" in developing a comprehensive computer-based environment for survey research. During the next several years, research and development projects will explore a variety of approaches to coordinating or integrating information-processing activities across the major stages of the survey research process. All such efforts can be associated with one of three basic strategies, with fairly obvious differences in costs and risks:

- Select a single system for each major phase and build linkages or translation programs to move information from each specific system (and stage) to the others (a *one-to-one* strategy for integration);
- Develop a single *comprehensive system* for all stages of the research process (a goal whose

scope has sometimes been described as analogous to an "airlines reservation system" for survey research);

- Develop general-purpose or system-neutral procedures for *data description*, so that users could move data from one system to any other system that uses the same external structure for data description (a *many-to-many* strategy for integration).

The high cost of the first of these strategies has already been discussed above. Until some other plan is successful, however, bilateral (one-to-one) linkages will continue to be the only solution.

At this point, it seems highly unlikely that any project or group will be successful in pursuing the second strategy, i.e., developing a single system which offers all of the services required for the collection, analysis, and management of survey-related information. Current systems in each area are based on internal structures that will be very difficult (or time-consuming) to reproduce within a system that also covers the other stages. Despite those obstacles, a comprehensive system for all survey-related activities represents an important long-term challenge, so that some researchers should seek the (substantial) resources that will be required to *design* such a package. If and when such a design is completed, it should be carefully evaluated before any actual development takes place, for the costs associated with a false start could be extremely large. In the meantime, however, it seems safe to assume that no comprehensive system will soon emerge which covers all three aspects or stages of the survey research process.

3.4. Cross-system integration through data description

That assumption, coupled with the recurrent need for transferring data between alternative systems for data collection, management, and analysis, suggests that our field might benefit from a common (or neutral) standard for storing and documenting the data produced by survey procedures. In such an approach, all cooperating systems for collection, management, or analysis would accept input data (and documentation) that have been stored in a common (or standard) format and could produce output data (and documentation) in that same format. Each such system would then only need to convert data to and from that common (neutral) format, rather than developing a different conversion program for each other system. The survey field has not agreed on a format standard for "system-neutral" data description, but discussions take place from time to time concerning potential alternatives.

As a first step in this direction, the CSM Program is developing a Data Description Language (DDL) which could become a common format for transmitting survey data and documentation between the several types of systems discussed above. Aspects of the intended language are already used to describe input and output data for CSM programs for Conversational Survey Analysis (CSA),[9] and procedures are now being completed to automatically convert data and Q instruments from a CASES project to the same (DDL) format, and to automatically generate setups from DDL for other statistical packages, (e.g., SPSS and SAS). This approach to data conversion between systems is summarized in Fig. 4.

The current DDL includes only those data elements required for CSA, but a comprehensive language of this sort should include documentation for survey-based information produced by all of the activities described at the beginning of this essay. Several survey organizations (including NASS) are considering a "database ap-

[9] The objectives and current status of Conversational Survey Analysis are beyond the scope of this paper. For a brief discussion of CSA's design and capabilities, see the CSA User's Guide (CSM Staff, 1989).

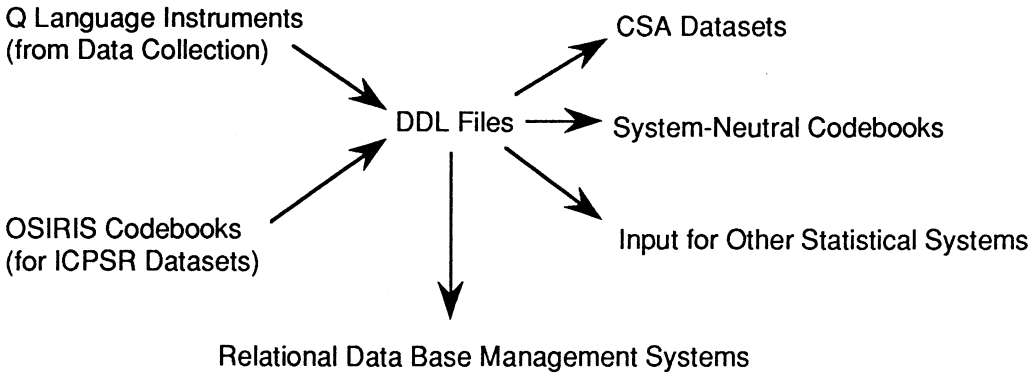


Fig 4. A "Neutral" Format for Transferring Data Between Systems Sources and Uses for DDL Files

proach" to managing all of their surveys, but it will be some time before specifications are completed for all of the data elements and structures for a demonstration project of that sort. See Tortora, Vogel, and Shanks (1985).

3.5. Alternative approaches to integration: Combining data collection and management

The above approach – based on a standard format for data description – represents only one of several strategies for overcoming the incompatibilities between alternative systems for data collection, analysis, and management. As suggested above, at least one group of specialists (in surveys and computing) should begin the process of designing a single (or comprehensive) system which provides all of the services required. A variant on that approach is now receiving increasing attention within organizations that must manage large and complex collections of survey data. In that approach, a relational database management system provides a common database environment (and computational capabilities) within which other programs can be accessed for either data collection or analysis.

Specifically, survey specialists from several organizations have advocated that data collection packages be revised to read and write all of their files in the internal format required by ORACLE (or SYBASE, INGRES, DBII, etc.), so that these study-level files can be integrated with other types of information and large collections of data sets. This strategy may be represented by the following (quite different) relationship between the various stages in the survey process:

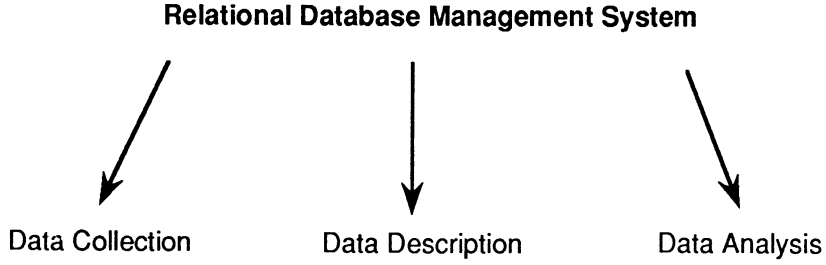


Fig. 5. Relational Database Management System

In such a design, the dominant status of a database management system should not be mandatory, for users must still be able to combine or link existing programs for collection, description, and analysis – as described above. Several organizations, however, are experimenting with a comprehensive "database approach" of this sort, in which all survey information is managed by a single relational system. NORC is already working on such a system (private communication with R. Baker 1989), and CSM is discussing a "relational" version of CASES with NASS and BLS.

This essay is neutral with respect to the comparative difficulty or long-term effectiveness of these alternative strategies, and encourages a variety of organizations to at least design the projects that will be required to answer such questions. No matter which approach emerges as the most effective strategy for integrating the survey process, however, the concepts and techniques used in RDBM systems are likely to play an important role. This article agrees with "database" theorists who have argued that complex data structures are best handled by decomposing those structures into a series of simpler (rectangular) files, and by representing the complexity involved in terms of relationships between those files. That perspective suggests that survey researchers will be more successful in integrating their diverse computer-related activities if all of their survey information is represented by (multiple) rectangular files and relationships (or linkages) between files. Development projects that deviate from this principle should be less successful in the long run, and this expectation should become increasingly important as the survey activities involved become more complex and comprehensive.

3.6. *The impact of new technology*

Since the late 1970s, many survey organizations have changed their internal division of labor in response to new systems for computer-assisted data collection. As emphasized in previous sections of this article those changes have involved a steady increase in the concentration of technical responsibility, as computer-based "instruments" specify more and more of the survey procedures involved. Thus, the interview schedule in a computer-assisted survey often defines sample elements and outcomes, and the same instrument may be augmented to control post-interview data entry or revision and document the resulting data. The substantial overlap in machine-readable information between stages and activities, coupled with improved procedures for transferring data between different systems (for data collection, analysis and management) have encouraged the integration of all computer-related activities for a given survey under a single study director. After more than a decade of experimentation and development, this trend toward computer-based integration (of previously separate survey activities) is now well underway, and recent developments in computing technology will almost certainly facilitate that process.

In particular, survey professionals are now discovering the possibility of visual integration of previously separate activities, based on high performance workstations. In this new computing environment, individuals who are responsible for coordinating survey activities across previously separate stages can control those tasks as simultaneous "windows" on a large screen attached to a computer that resides on their desk. These workstations offer a noticeably large display (at least 19 inch) with much greater resolution (more than a million pixels), a larger memory (several million bytes), and much more computing power (several million instructions per second) than in the environment that most of us have used until quite recently. These capabilities will almost certainly accelerate the integration of activities previously carried out by separate individuals, for a single study director can now move quickly

between windows in which individuals can carry out a variety of simultaneous operations. This kind of "workstation integration" is suggested in the screen shown on the following page, which contains a separate window for questionnaire administration, questionnaire development (or modification), statistical analysis, and identifying cases with specific characteristics.

The technical integration of survey activities is also being encouraged by the changes now taking place in communications between computers. Until fairly recently, survey information could only be processed on a given computer by physically moving all of the files to that system. With the growth of high-speed networks and distributed file systems, however, survey researchers can use their own (local) computer to process information that is stored on different (remote) machines. Survey researchers will soon be using high-speed communications and distributed file systems in a variety of contexts, including:

- immediate access to large databases from previous surveys, for re-analysis or comparison with current results;
- rapid movement of information between geographically separate systems for data collection and analysis;
- use of inexpensive workstations for data collection, so that larger computers in the same local network can be dedicated to data storage and retrieval; and
- "online" access to other computers during data collection, for circumstances in which information must be retrieved from large (external) databases before determining the next appropriate question.

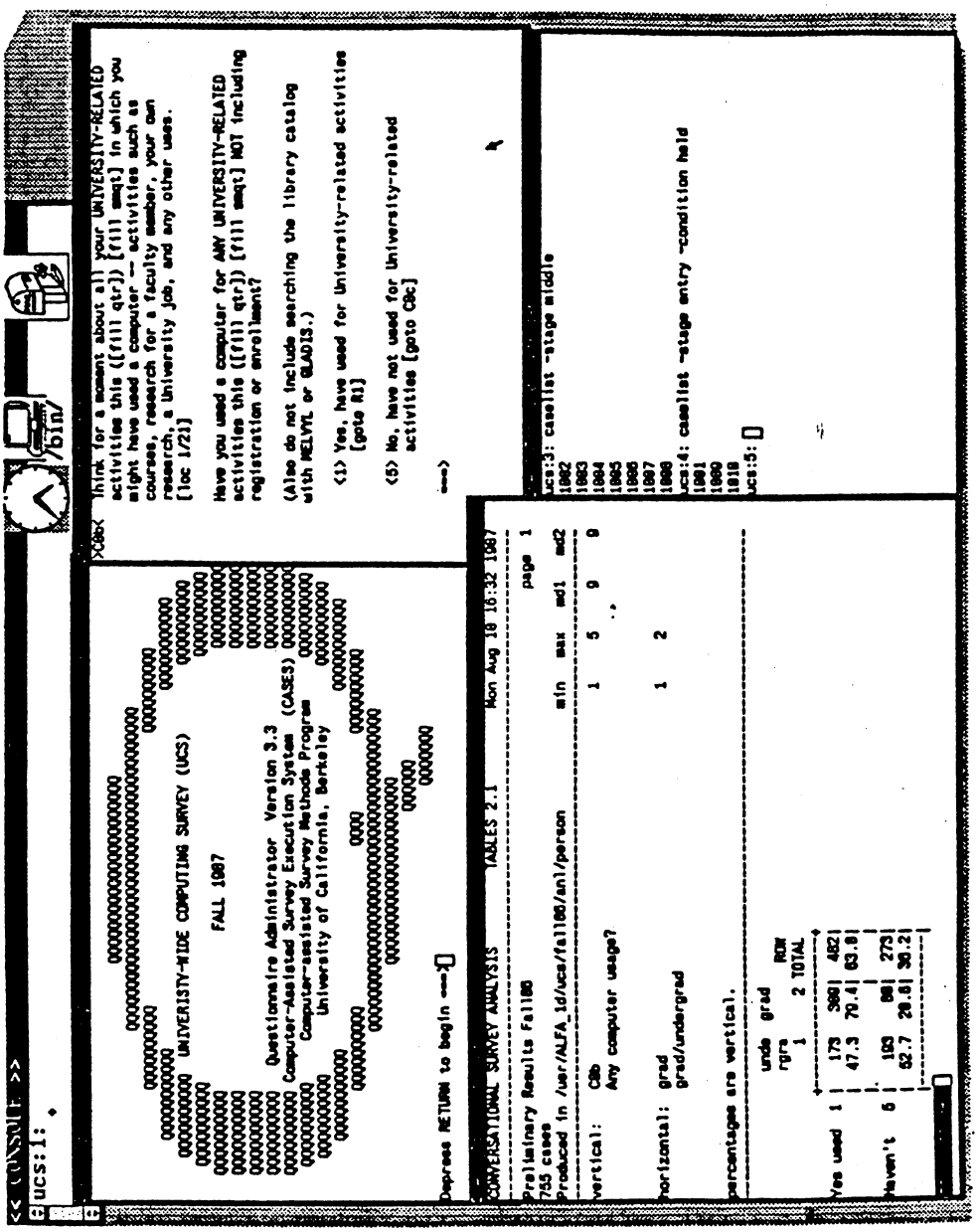
As emphasized in previous sections of this essay, the concept of a "structured questionnaire" is already changing, because of the personal computer's new capabilities for input and output (including images and sound). When coupled with advanced function workstations and highspeed communications, those capabilities will also contribute to the general trend toward survey integration, for new forms of data collection will present corresponding challenges for data management and documentation.

3.7 Summary

During the next few years, research and development in the survey field will continue to involve the improvement and generalization of separate systems for data collection, analysis, management, and documentation. Much remains to be done in each of these areas in order to take advantage of the developments now taking place in information technology, including workstations, graphics, large scale databases, and high-speed communications. In addition to these new capabilities, however, survey researchers will be exploring alternative approaches to integrating the entire research process. The so-called "database approach" has been enormously successful in providing integrated systems for collecting, managing, and displaying information in many other fields. The process of collecting survey-type data, however, is different from the activities handled by existing database management systems, and our combined requirements for "information processing" exceed those likely to be provided by current systems for data collection, analysis, or management.

As a consequence, survey researchers are now designing, developing, and testing a variety of ways to combine or coordinate their use of all three types of systems. At this point, we can only speculate about the comparative merit of those approaches. Hopefully, a sequel to this essay will identify strategies which have been successful in providing a unified (and simplified) approach to "information processing" in survey research.

THE SURVEY RESEARCHER'S WORKSTATION



4. References

Baker, R. P. (1987): Information Systems in Survey Research. Proceedings of the Bureau of the Census Third Annual Research Conference, Baltimore, Maryland, March 29–April 1, pp. 166–177.

Codd, E. F. (1970): A Relational Model of Data for Large Shared Data Banks. Communications of ACM, 13(6).

CSM Staff (1987): User's Guide to CASES; The Computer-Assisted Survey Execution System. University of California, Berkeley, Computer-assisted Survey Methods Program (CSM).

CSM Staff (1988): User's Guide to Conversational Survey Analysis (CSA). University of California, Berkeley, Computer-assisted Survey Methods Program (CSM).

Denteneer, D., Bethlehem, J. G., Hundepool, A. J., and Keller, W. J. (1978): The BLAISE System for Computer-assisted Survey Processing. Proceedings of the Bureau of the Census Third Annual Research Conference, Baltimore, Maryland, March 29–April 1, pp. 112–127.

Francis, I. (1981): Statistical Software: A Comparative Review. Elsevier, North Holland.

Freeman, H. E. and Shanks, J. M., eds. (1983): The Emergence of Computer-assisted Survey Research. Sociological Methods and Research, 12(2), pp. 115–118.

Nicholls, W. L., II and Groves, R. M. (1986a): The Status of Computer-assisted Telephone Interviewing: Part I – Introduction and Impact on Cost and Timeliness of Survey Data. Journal of Official Statistics, 2(2), pp. 93–115.

Nicholls, W. L., II and Groves, R. M. (1986b): The Status of Computer-assisted Telephone Interviewing: Part II – Data Quality Issues. Journal of Official Statistics, 2(2), pp. 117–134.

Palit, C. and Sharp, H. (1983): Microcomputer-assisted Telephone Interviewing. Sociological Methods and Research, 12(2), pp. 169–189.

Raskins, R. (1989): Statistical Software for the PC: Testing for Significance. PC Magazine, 8(5), March 14, pp. 103–116.

Shanks, J. M. (1983): The Current Status of Computer-assisted Telephone Interviewing: Recent Progress and Future Prospects. Sociological Methods and Research, 12(2), pp. 119–142.

Shanks, J. M. and Tortora, R. D. (1985): Beyond CATI: Generalized and Distributed Systems for Computer-assisted Surveys. Proceedings of the First Annual Research Conference of the Bureau of the Census, (March).

Sonquist, J. (1977): Computing and Surveys. Prentice Hall, Englewood Cliffs, New Jersey.

Statistics Sweden (1989): Computer-assisted Data Collection in the Labour Force Survey. A Report of Some Technical Tests. Technical report.

Tortora, R. D. (1984): CATI in an Agricultural Statistics Agency. U. S. Department of Agriculture, Statistical Reporting Service.

Tortora, R. D., Vogel, F. A., and Shanks, J. M. (1985): Computer-Aided Survey Methods. U. S. Department of Agriculture, Statistical Reporting Service.

Werking, G., Tupek, A., and Clayton, R. (1988): CATI and Touchtone Self-Response Applications for Establishment Surveys. U. S. Bureau of Labor Statistics. Paper presented at the Fourth Annual Census Bureau Research Conference, Arlington, Virginia (March 20–23).

Received August 1988
Revised April 1989

Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 23–39

The Optimal Design of Quality Control Samples to Detect Interviewer Cheating

Paul P. Biemer and S. Lynne Stokes[1]

"He is not cheated who knows he is being cheated."
Sir Edward Coke, Institutes (1628)

Abstract: Without interviewer quality control, interviewer cheating can seriously affect the accuracy of survey results. This paper proposes a method for designing quality control samples which maximizes the probability of detecting cheating for a fixed cost. First, data on interviewer cheating from a recent U.S. Bureau of the Census study are presented. Then a statistical model for describing dishonest interviewer behavior is proposed which assumes cheating is a random event governed by a probability distribution whose parameters depend on the interviewer. These parameters control the frequency and intensity of cheating as well as the geographic clustering of the falsified units.

A general quality control sample design and several associated cost models are proposed. A procedure for optimally choosing the sample design parameters according to specific types of interviewer behavior is described. Finally, the procedure is applied to optimize the interviewer quality control system used by the U.S. Bureau of the Census for the Current Population Survey and other current surveys.

Key words: Current Population Survey; National Crime Survey; reinterview; "curbstoning;" nonsampling error; survey costs.

1. Introduction

Every survey data collection organization, especially those that conduct personal interviews, must deal with the problem of interviewer cheating. The most blatant example of cheating occurs when an interviewer fabricates the responses for an entire questionnaire. Sometimes, however, cheating takes a more subtle form. For example, an interviewer may ask some questions in an interview and fabricate the responses to others. An interviewer may deliberately deviate from prescribed procedures, such as conducting a telephone interview where a face to face interview was indicated or conducting the interview with a willing but inappropriate respondent.

One of the most common methods used for detecting interviewer cheating in personal interview surveys is the verification method. For

[1]Paul Biemer is Head, Department of Experimental Statistics, and Director, University Statistics Center, New Mexico State University, Las Cruces, NM 88003-3130, U.S.A. S. Lynne Stokes is Assistant Professor, Department of Management Science and Information Systems and Center for Statistical Sciences, University of Texas, Austin, TX 78712, U.S.A.

this method, a sample of an interviewer's assignment is recontacted in order to verify that an interview was conducted as required and that (at least) the critical components of the questionnaire were obtained accurately.

The question we address in this paper is how to design the verification sample in order to maximize the probability of detection of a cheating interviewer at least once during a specified time period. The methodology developed is appropriate for organizations whose interviewer staff is stable and whose interviewers participate regularly in surveys in which they have similar workloads. Although the emphasis here is on "in-the-field" interviewing (i.e., face to face and decentralized telephone interviewing), the methodology is adaptable to centralized telephone interviewing. In that case verification takes the form of a system of unobtrusive telephone monitoring.

Since the resources allocated to this aspect of a survey's quality control program is generally quite limited, only a small portion of the interviewer's workloads can be verified. The competing choices we allow to be made concerning the verification design are (a) how often the interviewer is chosen for verification (b) how much of his/her assignment is inspected when he/she is chosen and (c) what size the sampling units (persons, households, or groups of households) should be. The optimal choice depends on an individual interviewer's cheating behavior and the cost of the design choices.

In Section 2 we review information which has appeared in the literature concerning interviewer cheating behavior. In addition, the data resulting from a program implemented by the U.S. Bureau of the Census in 1982 to collect such data are reported. In Section 3, a model for interviewer cheating behavior is suggested and the probability of detection for a given verification scheme is derived. Section 4 gives some empirical rules for optimal design for our cheating model for two typical cost function forms. Finally, in Section 5 the model is used to develop a verification sample design for the Current Population Survey (CPS), the largest demographic survey run by the U.S. Bureau of the Census. This application motivated the development of our model.

2. Interviewer Cheating Behavior

Interviewer cheating has long been recognized as a problem among survey organizations. Crespi (1945) conjectured that "almost every interviewer will eventually succumb [to cheating] ... if fabrication is made to appear the only practicable solution to the problems facing the interviewer." He suggested several factors, related to either the questionnaire or the administration of the survey, which may operate to demoralize the interviewer. Related to the questionnaire were: (1) questionnaire length and respondent burden, (2) poor questionnaire design, e.g., apparent repetition of the same questions, and (3) difficult or antagonistic questions. Among the administrative demoralizers are: (1) overly difficult assignments or inadequate remuneration, (2) improper or inadequate training, (3) use of part-time interviewers for whom the unpredictable demands of interviewing may compete with the necessities of another job and (4) external factors such as the weather, bad neighborhoods, roads, etc. which may operate to encourage cheating.

Crespi's proposed solution to the cheater problem is the dual strategy of (a) eliminating the demoralizers by careful and intelligent survey design and administration with ample opportunity for interviewer advisement (for example, the present-day "quality circles") and (b) using the verification method to deter cheating. Bennett (1948 a and b), Sheatsley (1951), Boyd and Westfall (1955), and Evans (1961) appeared subsequently which provided suggestions to help the survey practitioner implement part (a) of this strategy.

There is little guidance in the literature, however, on implementing (b). Making the prob-

lem worse is the paucity of published data on characteristics of interviewers who cheat or how they do it. This is understandable because cheaters are difficult to detect, and few studies have made the collection of such data their major goal. The first reported data was made for the National Opinion Research Center (Sheatsley (1951)) and included characteristics of the small number of interviewers who were dismissed for cheating between 1941 and 1949.

In 1982, the U.S. Bureau of the Census began a program to collect information on all confirmed or suspected cases of cheating by interviewers in their current surveys. The purpose of the data collection was to aid in the modeling of cheating behavior and ultimately in the selection of an optimal verification design. Cheating problems targeted were complete or partial fabrication of the survey responses as well as other improper interviewer conduct, such as use of proxy respondents in situations where self-response was required.

The first results available from this study covered the period September 1982 through August 1985 (Bureau of the Census (1986)). During that time, it was established that 140 interviewers (about 3–5 % of all interviewers) committed some form of cheating, and an additional 31 interviewers were suspected of cheating. Of the 140 confirmed cases, 100 were identified through the reinterview verification program. The remaining 40 cases were detected by other means, such as inspection of the returns, information provided by other interviewers, etc. Overall, most of the cheating (72 %) involved complete fabrication of interviews. The next most frequent violation was the misclassification of units as vacant when they were, in fact, occupied (17 %). Indeed, this form of cheating is just as damaging as falsifying an entire interview since the unit is then erroneously regarded as out-of-scope for the survey. In the National Crime Survey (NCS), which requires that each respondent answer for him/herself, 20 out of the 26 confirmed cases of cheating involved the violation of this self-response rule, often accompanied by other infractions as well.

Some further results of this study are now summarized.

1. For the two largest demographic surveys, the CPS and the NCS, 87 % of the falsified interviews occurred in urban areas, only 13 % in rural areas. Since roughly 70 % of the sample is located in urban areas for these surveys, there is evidence (statistically significant at the 5 % level of significance) of a higher degree of cheating in urban areas.
2. Table 1 shows the distribution of cheaters (CPS and NCS only) by years of service with the Bureau of the Census. Almost half of the confirmed violators have less than one year of service, while only 23 % of all interviewers have less than one year of service. These data indicate a substantial and highly significant tendency for relatively inexperienced interviewers to cheat more frequently than interviewers having one or more years of experience. Alternatively, the data may indicate an adeptness of more experienced interviewers for escaping detection of cheating.

Table 1. Distribution of CPS and NCS interviewers found cheating by years of experience

Length of service	Cheaters (%)	All interviewers (%)
Less than 1 year	46	23
1–2 years	13	28
3 years or more	43	48

3. Experienced interviewers (i.e., those with a year or more of experience) who were detected cheated at an average rate of 19 % of the households in their assignments for CPS and NCS. Furthermore, less than 13 % of these cheaters were involved in fabrication of responses. By contrast, interviewers with

less than one year of service displayed a tendency to cheat at a much higher rate, viz., an average of 30 % of the households in their assignments, with roughly half of the cheaters being involved with the complete fabrication of interviews.

3. The Model

In Section 3.1, a simple model is proposed that describes interviewer cheating behavior. This model views cheating as a random event governed by a specified probability distribution which depends on several parameters that control the frequency and pattern of cheating. In conjunction with this model, a general design for the verification sample is described. It is a variation of the one previously used by the U.S. Bureau of the Census for the CPS and is a generalization of the one in current use. This design is described in Section 3.2. In Section 3.3, the probability of detection of cheating by an interviewer for any specific verification design is derived. This probability will depend, of course, on the interviewer's cheating parameters. Finally, in Section 4.4, two models for the cost of verification sampling are proposed. If all parameters of the interviewer cheating and cost models were known, an optimal design could be selected. Such a design is defined to be one which maximizes the probability of detection.

One problem with this approach is that interviewers do not all behave alike, and therefore a design which is optimal for one may not be optimal for another. A solution to this problem is to divide the interviewers into strata defined by their frequency and pattern of cheating. For example, the data described in Section 2 suggest that interviewer experience would be a good stratifying variable for CPS interviewers. The optimal verification design parameters could be determined for each stratum.

The second problem with the approach, however, is that the parameters of the cheating model for individual interviewers or groups of interviewers are not known in advance. Furthermore, it is difficult to obtain enough information to estimate the parameters of the model proposed in Section 3.1. During the study described in Section 2, the interviewer was either fired or resigned in 97 % of the cases in which he or she was detected fabricating an interview. Information about patterns of cheating is not available when the behavior can be observed only once.

Nevertheless, the model developed can be useful for determining a verification design. Two possibilities for its use follow. First, one might optimize the design against the most damaging violators. This may be, for example, interviewers who falsify more than some specified fraction of their assignments. Second, one might use the model to choose a design, if one exists, which is nearly optimal against a wide range of likely interviewer behavior. The latter of these is the use made of the model for the CPS application described in Section 5.

3.1. *Model for interviewer cheating*

Consider a complex survey with any probability sampling design for which ultimate stage sampling units (USUs) are clusters of m interview units. For example, USUs may be geographical segments of m housing units or they may be households of m individuals. The time period for the survey during which interviewers are to be evaluated will be referred to as the observation cycle. Every interviewer is to be inspected at least once during an observation cycle. Let f denote the number of times the survey is repeated during one observation cycle and refer to these repetitions as interviewing periods.

Consider a particular interviewer for some interviewing period within an observation cycle. Let n denote the number of USUs in the interviewer's assignment and let the random variables $b_h\,(h = 0, \ldots, m)$ denote the number

of USUs having exactly h misrepresented (or fabricated) interview units. Hence $\sum_{h=0}^{m} b_h = n$.

In the next paragraph, we propose a probability distribution for $\mathbf{b}' = (b_0, \ldots, b_m)$ which is a mixture of two distribution functions: a Bernoulli distribution with parameter π and a multinomial distribution with parameter vector $\mathbf{p}' = (p_0, \ldots, p_m)$.

There are numerous factors which may influence an interviewer's decision to cheat during an interviewing period, as Crespi suggested. Some of these factors are always present, such as problems with a questionnaire. Others occur only occasionally, such as problems with the weather or interference from a part-time job. Consequently, some interviewers (those influenced by the ever-present problems) may be susceptible to cheating at all times. Others (those influenced only by irregularly occurring events) may cheat only when there is no other way to complete their assignments. We attempt to capture this behavior pattern in our model by defining a parameter π to be the probability that an interviewer considers or has some positive probability of cheating, and we refer to π as the frequency of cheating. Next we define $P(\mathbf{b})$ to be the probability distribution of $\mathbf{b}' = (b_0, \ldots, b_m)$ associated with the interviewer and assume that $P(\mathbf{b})$ is the multinomial distribution with parameters $\mathbf{p}' = (p_0, \ldots, p_m)$ and n. This multinomial assumption suggests that the probabilities of falsifying $h = 0, 1, \ldots, m$ units in a cluster is the same for every cluster. However, given that one or more units have been falsified in a cluster, the multinomial distribution allows us to change the probability that other units in that cluster will be falsified. That is, we may, through $\mathbf{p}$, arrange for a nonzero intracluster correlation for falsified units. (In the sequel, an interviewer whose cheating behavior is governed by the model with parameters $(\mathbf{p}, \pi)$ will be referred to as a $(\mathbf{p}, \pi)$ cheater.) Thus, p_h is the probability that h units in a given USU are misrepresented. We define $\bar{p}$, referred to as cheating intensity, to be the expected proportion of misrepresented interview units in a $(\mathbf{p}, \pi)$ cheater's assignment, given that he or she is susceptible to cheating; i.e.,

$$\bar{p} = \sum_{h=0}^{m} h p_h / m. \tag{3.1}$$

The decision to cheat for a particular unit may not be made independently of other units in an interviewer's assignment. For example, cheating may be concentrated within certain USU's which share characteristics that may influence the interviewer to cheat, such as nontelephone households, undesirable neighborhoods, or areas with difficult access. It is possible with this model to describe the strength of this clustering effect by appropriately defining the p_hs. If no clustering is present, for example, misrepresented units will be distributed among the interview units in each USU according to a binomial distribution, so that $p_h = \binom{m}{h}\bar{p}^h(1-\bar{p})^{m-h}$. Perfect clustering would mean that either all or no units in a USU are misrepresented; i.e., that $p_0 + p_m = 1$. One can show that these two extreme conditions yield extreme values (0 and 1 respectively) of the intracluster correlation

$$\delta = \mathrm{Cov}(y_{ij}, y_{ij'})/\mathrm{Var}(y_{ij}), \tag{3.2}$$

where $y_{ij} = 1$ if the jth unit in the ith USU is misrepresented and 0 otherwise. Correct specification of the magnitude of δ is important, since it has an impact on the optimal reinterview sample design.

This simple model can describe a wide range of interviewer cheating behavior. The consistent, low-level cheater can be modeled by setting π large and $\bar{p}$ small, while the erratic cheater can be accomodated by setting π small. Clustering of the affected units can be modeled by the relative values of the p_hs.

Table 2. Notation

Symbol	Definition
n	number of USU's in an interviewer assignment
m	number of units in a USU
b_h	number of USUs having h falsified units; $\sum_h b_h = n$.
I	number of interviewers for the survey
f	number of interviewing periods in an observation cycle
s	number of interviewers selected for each supplementary sample
ℓ	number of USU's to be reinterviewed in each interviewer assignment
t	number of units to be reinterviewed in each sample USU
π	probability that an interviewer is susceptible to cheating
p_h	probability h units in a USU are falsified
δ	intracluster correlation coefficient for cheating; see (3.2).

3.2. *The verification sample design*

One major goal of the verification sample is to detect interviewer cheating; a second is to deter it. In order to meet the first goal, we search for a design which, for a given cost, will maximize the probability of selecting the units in a sample which are misrepresented. So that the design we choose will be a deterrent, we require the following:

1. Every interviewer must be selected at least once during an observation cycle.
2. The selection of interviewers and units must be unpredictable by the interviewers.

Many designs satisfy these criteria. One of them, which is now used by the U.S. Bureau of the Census, will illustrate the design optimization methodology. For simplicity, we assume a single population of interviewers having cheater parameters $(\mathbf{p}, \pi)$. For the case where interviewers are divided into strata, the procedure described below may be applied separately in each stratum.

Let the I interviewers for the survey be divided into f mutually exclusive and exhaustive groups. (Recall that f is the number of interview periods in an observation cycle.) To simplify the subsequent formulas, we assume that I is evenly divisible by f. Randomly order the groups. Then all interviewers in group i will be selected for verification in the ith interviewing period. This group will be referred to as the ith predesignated sample. In addition, let a specified number, say s, of interviewers be selected at random from the remaining $I(1-1/f)$ interviewers. We refer to this group as the supplementary sample. This group introduces more unpredictability into the selection of interviewers. From each selected interviewer's assignment a sample of ℓ USU's is randomly chosen. Within each USU, t units are reinterviewed. Thus, the total sample size in each interviewing period is $(I/f+s)\ell t$.

If cheating is observed for any of the ℓt interview units in an interviewer's assignment, we say that a cheater was detected. Our objective is to find the verification design parameters s, ℓ, and t which maximize the probability of detecting a cheater for a fixed total cost under a specified interviewer cheating model.

3.3. *Derivation of detection probability*

The probability of detecting a cheater with verification design parameters, (s, ℓ, t) will be denoted by $D(s, \ell, t)$. In Appendix 1, it is shown that, for a $(\mathbf{p}, \pi)$ cheater

$$D(s, \ell, t) = 1-[1-P_t(\mathbf{p}, \pi)]\left[1-\frac{fs}{I(f-1)}\right]P_t(\mathbf{p}, \pi)^{f-1}, \quad (3.3)$$

where

$$P_t(\mathbf{p}, \pi) = \pi(1-\eta_t^{\ell}) \text{ and } \eta_t = \binom{m}{t}^{-1} \sum_{h=0}^{m} \binom{m-h}{t} p_h, \quad (3.4)$$

where $\binom{m-h}{t} = 0$ if $t > m-h$. η_t can be described as the probability that a $(\mathbf{p}, \pi)$ interviewer is not observed cheating in a sampled USU during an interviewing period in which he is susceptible. $P_t(\mathbf{p}, \pi)$ is then the conditional probability that a $(\mathbf{p}, \pi)$ cheater is observed cheating in the ith interviewing period, given that he or she is selected in either the ith predesignated or corresponding supplementary samples. (Note that $P_t(\mathbf{p}, \pi)$ does not depend upon i.)

One special case of (3.4) that will be considered in detail is when all interview units are inspected within the sample USU's, or $t = m$. In that case $P_t(\mathbf{p}, \pi)$ becomes

$$P_m(p_0, \pi) = \pi(1 - p_0^{\ell}). \tag{3.5}$$

Then only π and p_0 need be specified to compute $D(s, \ell, t)$.

3.4. A general cost model for design optimization

The cost model described in this section is general enough to apply for many verification operations. Let $K(s, \ell, t)$ denote the variable costs for an observation cycle of a design having parameters s, ℓ, and t. Then, for constants C_1, C_2, C_3, and C_4,

$$K(s, \ell, t) = (I + fs)[C_1\ell + C_2\ell t + C_3(t)\sqrt{\ell} + C_4] \tag{3.6}$$

where C_1 is the cost associated with each sampled USU's (for example, a sampling cost); C_2 is the cost incurred for each unit inspected; $C_3(t)$ allows for a travel cost for situations in which interviewers must travel varying distances to the USU's. This cost will depend upon t if the cost of travel to inspect all units in a USU depends on the number of units verified in a USU. C_4 is the fixed cost associated with each sampled interviewer.

Two special cases of (3.6) will be considered in Section 4. The first is the case where cost is simply proportional to the total sample size, i.e., $C_1 = C_3 = C_4 = 0$, or

$$K_1(s, \ell, t) = (I + fs)C_2\ell t. \tag{3.7}$$

This cost function might be appropriate for a verification program which relies solely on telephone reinterviewing of a sample of survey respondents. There are no costs for travel or any other disaggregate costs associated with the number of sample USU's.

For designs in which all the interview units are revisited in person, but moving from one unit to another within a USU does not incur much additional travel costs, a reasonable cost function may be obtained setting $C_1 = 0$ in (3.6) and allowing $C_3(t) = C_3$. Then

$$K_2(s, \ell, t) = (I + fs)\ (C_2\ell t + C_3\sqrt{\ell} + C_4). \tag{3.8}$$

The cost of traveling to the ℓ USUs is proportional to $\sqrt{\ell}$ if the USUs are randomly distributed within the interviewer's assignment area (Pielou (1969, p. 111)).

4. Some Rules for Optimal Design

In this section, we state some general rules, some analytical and some empirical, concerning the optimal choice of a verification design when cost functions of the form given by (3.7) or (3.8) are appropriate. The optimal design is determined by maximizing the detection probability $D(s, \ell, t)$, subject to the constraint $K(s, \ell, t) = C_F$, where C_F is the total fixed cost for the verification sample.

Rules 1 through 3 deal with determining an optimal trade-off between the frequency that an interviewer is sampled (s) and the thoroughness of the inspection of his or her assignment (ℓ). The number to be selected from each USU is assumed fixed, which is equivalent to assuming a fixed value for η_t, defined in (3.4).

Rule 1. Assume the cost function $K_1(s, \ell, t)$ in (3.7) and $\pi = 1$. Then for t held fixed, any values of s and ℓ satisfying $K_1(s, \ell, t) = C_F$ will provide a sample design that is either optimal or near optimal.

Rule 1 is supported by Theorem 1, which is stated and proved in Appendix 2. Theorem 1 says that, under the conditions of Rule 1, the maximum detection probability for fixed cost can be achieved by choosing s at either of its extremes (i.e., $s = 0$ or $s = \frac{f-1}{f}I$). Numerical investigations suggest further that the detection function $D(s, \ell, t)$ is very flat over the entire range of s for likely choices of C_F and $\mathbf{p}$.

Therefore, when the cost model described in (3.7) is appropriate, the probability of detection of the consistent cheater varies little with the choice of s and ℓ. The verification design choice can safely be made, then, on the basis of other considerations. For example, a design which is optimal for erratic cheaters could be implemented, and the survey managers could be assured that it would be near optimal for the consistent ones as well.

Rule 2. Assume the cost function $K_2(s, \ell, t)$ in (3.8) and $\pi = 1$. Then for t held fixed, choosing s small but non-zero such that $K_2(s, \ell, t) = C_F$ will provide an optimal or near optimal design.

Rule 2 is supported by Theorem 2, which is stated and proved in Appendix 2. Theorem 2 says that, under the conditions of Rule 2, the probability of detection obtained by choosing s at its minimum ($s = 0$) is always larger than that obtained by choosing s at its maximum ($s = \frac{f-1}{f}I$). Numerical investigations further suggest that a choice of $s = 0$ actually maximizes the detection probability for likely choices of $\mathbf{p}$, C_F and for a wide range of cost function parameters C_2, C_3, and C_4.

In practice, however, the choice of no supplementary sample ($s = 0$) eliminates the unpredictability of the verification design. Since unpredictability was one of its specifications, this choice is not acceptable. The best strategy, as Rule 2 suggests, is therefore to choose s small, but non-zero.

When $\pi < 1$, numerical investigations have shown that no verification design exists which will come close to maximizing the probability of detection over all reasonable specifications of the cheating and cost models for either form of the cost function. Examples can be found for which detection probability is maximized at either extreme or at intermediate values of s. Furthermore, the loss in detection probability from a poor choice of a survey design is sometimes large. This observation suggests the following rule.

Rule 3. No optimal verification design exists for detecting an interviewer having $\pi = \pi_0 < 1$. Instead, the best strategy is sensitive to $\mathbf{p}$ and parameters of the cost models. Therefore, the best possible information about these parameters should be collected and sensitivity analyses performed to aid in the choice of design.

An application of Rule 3 is illustrated in Section 5.

Our discussion so far has dealt solely with the choice of s and ℓ. Now we turn to the problem of determining the optimal choice for t. It is affected by the magnitude of δ, defined in (3.2). Rules 4 and 5 address the choice of t when δ assumes one of its extreme values.

Rule 4. If there is no clustering of misrepresented units (i.e., $\delta = 0$) and the cost function is $K_1(s, \ell, t)$ defined in (3.7), then any choice of t is equally good. However, if the cost of sampling a new USU is greater than the cost of sampling a comparable number of units

within existing sample USUs, as for $K_2(s, \ell, t)$, then $t = m$ is optimal.

Rule 5. If misrepresented units are perfectly clustered (i.e., $\delta = 1$), then the best strategy is to choose $t = 1$ regardless of the cost function.

Rules 4 and 5 follow from the observations about δ made in Section 3. If there is no clustering within a USU, an identical amount of information can be obtained from sampling two units within the same USU as from different USU's. Therefore the optimal choice depends on which is cheaper. Since for $K_1(s, \ell, t)$, either choice has identical cost, the choice of t is unimportant. For $K_2(s, \ell, t)$, there is a saving of costs associated with remaining within a USU, so the best strategy is to choose t as large as possible. When there is perfect clustering of misrepresented units, selecting more than one unit per USU buys you no information and thus is wasteful, if it costs anything at all.

For those frequent cases in which δ is between the two extremes, the optimal choice of t is not so easy to make. However, the two rules together suggest that when cost is directly proportional to the number of units verified (as for K_1), $t = 1$ should be chosen. When K_2 is the appropriate cost function, an analysis such as that undertaken in Section 5 is required to determine the optimal t.

5. An Application to the Current Population Survey

5.1. Description of the Current Population Survey

The CPS provides the official labor force statistics for the United States. The survey is conducted monthly by the U.S. Bureau of the Census and has consisted of between 50 000 and 60 000 household interviews per month. In addition to data on employment and unemployment, the survey also provides information on annual household and individual income.

The sample is a stratified multistage cluster sample having USU's which are typically clusters of four neighboring households. The segments are selected at random within primary units which are essentially counties or groups of counties. Interviewer assignments generally average about 12 segments or about 48 housing units. Due to the rotational design of the sample, about an eighth of the housing units in an assignment are new to the program, an eighth are being interviewed for the second time, and so on up to an eighth being interviewed for the eighth and last time. Between 30 % and 40 % of the interviews are conducted face to face while the remainder are conducted by telephone.

The CPS interview quality control program consists of three components: (a) a reinterview survey to detect interviewer cheating, (b) an inspection of all the interview forms by clerks who are specially trained to detect interviewer errors in completing the forms, and (c) an annual on-site observation of the interviewer by a supervisory representative as the interviewer completes an assignment. The purpose of the annual observation is to provide an expert evaluation of the interviewer's interviewing technique. The remainder of the section will be concerned with the sample design of the quality control reinterview survey.

5.2. The quality control reinterview

For a sample of households, the reinterviewer, who is typically a senior interviewer and supervisory representative, re-asks some or all of the questions on the original questionnaire. Any discrepancies between the original interview and the reinterview are reconciled with the respondent. The reinterviewer also determines whether the discrepancy was the fault of the interviewer or the respondent. In addition to detecting interviewer cheating, the reinterview also serves as a device for detecting both deliberate and unintentional errors that occurred in

the original interview. The number of errors found in the interviewer's assignment that are the fault of the interviewer are tallied and these results are immediately reported to the interviewer for corrective action. Whenever possible, the reinterviews are conducted by telephone. Otherwise, a face to face reinterview is conducted and travel costs are incurred.

Prior to 1982, the CPS quality control reinterview program required that each interviewer be randomly selected for reinterview twice per year, once during the first six months of the year and again in the last six months. Each time an interviewer was selected, all the households in a randomly chosen third of the approximately 12 USU's in his/her assignments were reinterviewed. Thus, about one eighteenth of all CPS households (i.e., one sixth of the interviewers and one-third of each interviewer's assignment) were reinterviewed each month. This design was flawed since the time of reinterview was somewhat predictable. For example, an interviewer selected for reinterview in January could not be selected again until July. Many interviewers were aware of this pattern so that the reinterview was less effective as a deterrent.

In 1982, a redesigned CPS quality control reinterview program was implemented to correct this deficiency. Prior to implementation a study was conducted to aid in the selection of an improved sample design (Biemer, Judkins, Schreiner, and Stokes (1982)). The sampling scheme described in Section 3 was adopted for the CPS and subsequently for all the Bureau's continuing demographic surveys. The observation cycle was chosen to be twelve months, so that the predesignated sample consisted of $I/12$ interviewers. The design options to be determined, then, were the number of USU's to be sampled from each interviewer's assignment (ℓ) and the number of households to sample from each USU (t). The size of the supplementary sample (s) was determined by the cost constraint. We begin with a description of the cost function for the reinterview program.

5.3. *Cost function*

The cost of reinterviewing an interviewer's assignment can be decomposed into several components. There will be costs associated with the time involved in conducting face to face reinterviews and different costs for telephone reinterviews. Other costs arise from reconciling differences between interview and reinterview responses, completing the reinterview forms, and the time involved in discussing the results with the interviewers. For face to face reinterviews, mileage costs and cost for reinterviewer time while traveling and while conducting the reinterviews will be incurred. For telephone reinterviews, there will be no travel costs; however, telephone toll charges may be incurred.

A detailed analysis of the CPS reinterview costs was conducted and it was determined to be well-described by (3.6), but with $C_1 = 0$. Then the cost coefficients C_2, $C_3(t)$, and C_4 were estimated. The details of this analysis are documented in a Bureau of the Census report (Biemer et al. (1982)). In addition to a national cost function, a number of subnational cost functions were developed corresponding to urban, suburban, and rural areas where travel costs differ substantially. For the present illustration, only the national model will be considered. For a reinterview survey with design parameters (s, ℓ, t), the annual variable cost is given by the following model:

$$K(s, \ell, t) = (I + 12s)(2.41\,\ell t + C_3(t)\sqrt{\ell} + 79.58), \quad (5.1)$$

where

$$C_3(t) = 7.19\sqrt{1 - .85^t}. \quad (5.2)$$

The choice of $C_3(t)$ is explained by noting that the per USU cost depends on the average number of visits required for each USU in the reinterview sample, information which was not directly available. However, it may be assumed that a visit to a USU was made only if one of the t sampled households within it required a

personal visit. Since roughly 85 % of the CPS reinterview households have face to face reinterviews, we estimated by assuming a binomial distribution that approximately $(1-.85^t)\ell$ USUs out of the ℓ sampled USUs would contain at least one such household. Thus, travel costs increase in proportion to the square root of this expression (see comment under (3.8)). Data from the main CPS survey were used to estimate the travel costs associated with each segment visited. The coefficient 7.19 in (5.2) includes the cost for travel time as well as mileage.

The coefficient $C_2 = 2.41$ is an average cost for reinterview time associated with each sample unit. C_2 includes the cost of the time spent for a face to face reinterview and the cost of the time for a telephone reinterview averaged over all reinterviewer assignments. Finally, $C_4 = 79.58$ is the average cost for reinterviewer travel from his/her home base to the area of the interviewer's assignment.

The optimization results which follow were found, through sensitivity analyses, to be quite robust to absolute errors in these coefficients. More critical are the relative errors among the coefficients, i.e., the proportion of total costs accounted for by each cost component. Yet even here quite substantial changes in the relative sizes of C_2, $C_3(t)$, and C_4 had only moderate impact on the optimal design.

5.4. Determining the "near" optimal design of the CPS

Since virtually nothing was known at the outset about interviewer cheating behavior, it was not possible to select an optimal reinterview design with the model described in Section 3. Instead, the model was first used to determine to what extent the detection probability was sensitive to the design choice over likely ranges for the parameters $\mathbf{p}$ and π. Rule 2 of Section 4 suggests that the trade-off between s and ℓ is not critical for consistent interviewers ($\pi = 1$). However, it was believed that the conditions causing cheating by CPS interviewers were temporary, and thus erratic cheaters ($\pi < 1$) should be considered the target. Rule 3 says that in this case, the optimal choice of ℓ and s is highly variable. That knowledge led to the data collection program whose results are described in Section 2.

We did know from the start that misrepresented units were not perfectly clustered within USU's. What we did not know was if there was some or no clustering. Rules 4 and 5 of Section 4 tell us that if $\delta = 0$, the design already in use ($t = m = 4$) was best, but if even a slight amount of clustering is present, that design might be inefficient. Therefore we made sure that the data about cheaters was collected in such a way that δ could be estimated. Since such data would take years to amass, however, a program was also begun to collect data to allow estimates of correlation to be made for characteristics believed to be associated with misrepresented households, such as telephone ownership, income, and employment status.

After some information from these data collection efforts became available, the model was again used to aid in the selection of a reinterview design. The goal was to select a design whose loss in detection probability from that of the optimal choice would be small over the range of π and $\mathbf{p}$ we believed to be likely. A further goal was to compare this near optimal design with that of the design in use ($\ell = 4$, $t = 4$). This procedure is now described.

The data collected since 1982 gives a small amount of information about $\bar{p}$ and δ. From Section 2, recall that $\bar{p}$, the proportion of misrepresented units in a cheating interviewer's assignment, was observed to be 30 % or less. Therefore, we varied $\bar{p}$ in the interval $0 < \bar{p} \leq .30$. We also found that estimates of δ for characteristics believed to be associated with cheating were generally fairly small (less than .5), but non-zero. This led us to restrict our investigation to the values $0 \leq \delta \leq .5$. Fur-

ther practical considerations limited the design choice by restricting ℓ to the range $2 \leq \ell \leq 6$, since there was concern that sampling more than six (out of a possible 12) USU's in an interviewer's assignment might adversely affect the interviewer's cooperation rate in the assignment for the subsequent months of interviewing. Sampling less than two would not provide adequate work and compensation to employ a reinterviewer.

Within these constraints, a search was begun for a near-optimal design. The first choice made was that of t, for which the behavior of

$$M(\delta; t, \pi) = \max_{2 \leq \ell \leq 6} D(s, \ell, t)$$

was studied. Figure 1 illustrates with $\bar{p} = .10$ and $\pi = .1$, .5, and .9 the type of results obtained. In order to completely specify $\mathbf{p}$, p_2 and p_3 were taken to have values corresponding to completely random cheating and p_0, p_1, and p_4 were then chosen to satisfy the constraints stated above for $\bar{p}$ and δ, along with $\sum_j p_j = 1$. Except for $\pi = .9$, $t = 3$ appears to yield uniformly higher detection probabilities than $t = 4$, while $t = 4$ is preferred when π is .9. Note, however, that $M(t, \delta, \pi)$ increased less than 4 % for $t = 4$ relative to $t = 3$ in that case. When $\bar{p}$ was varied over the range (.05, .30), only the level of detection probability changed, not the choice of design parameters. Therefore, a choice of $t = 4$ was made since it was near optimal and maintained the status quo.

Since cost is fixed, only one other parameter, either s or ℓ, need be determined in order to completely specify the reinterview sample design. Furthermore, as can be noted from (3.5), the amount of clustering does not affect this choice when $t = 4$. In fact, only p_0 and π need be specified in order to completely determine the cheater behavior model when $t = 4$.

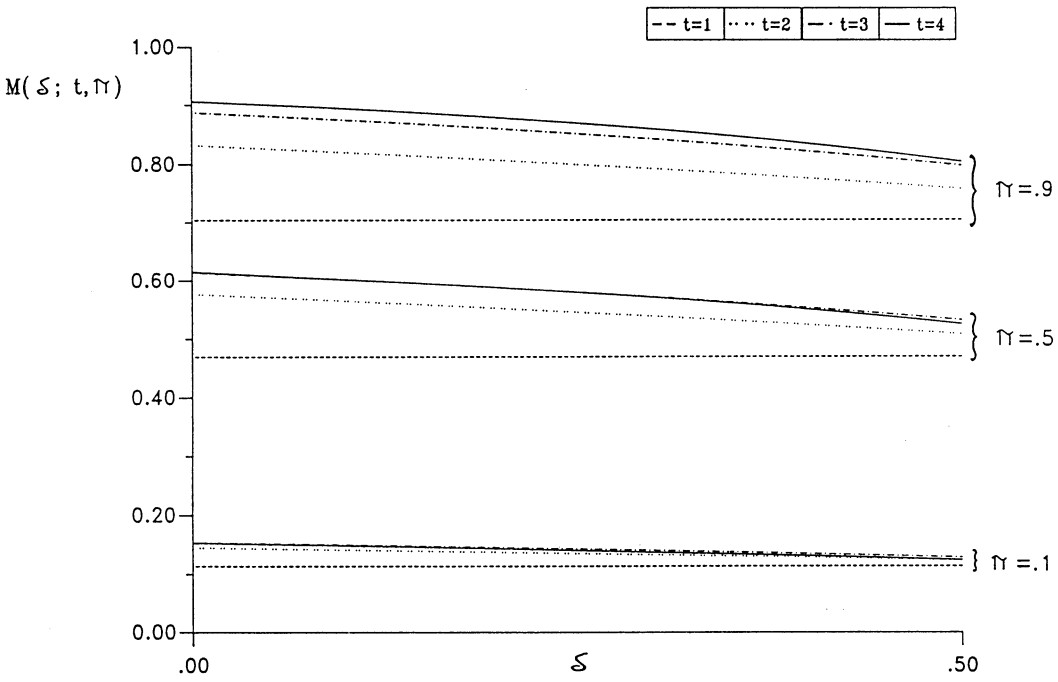


Fig.1. Maximum detection probabilities for given t as a function of δ for π = .1, .5 and .9

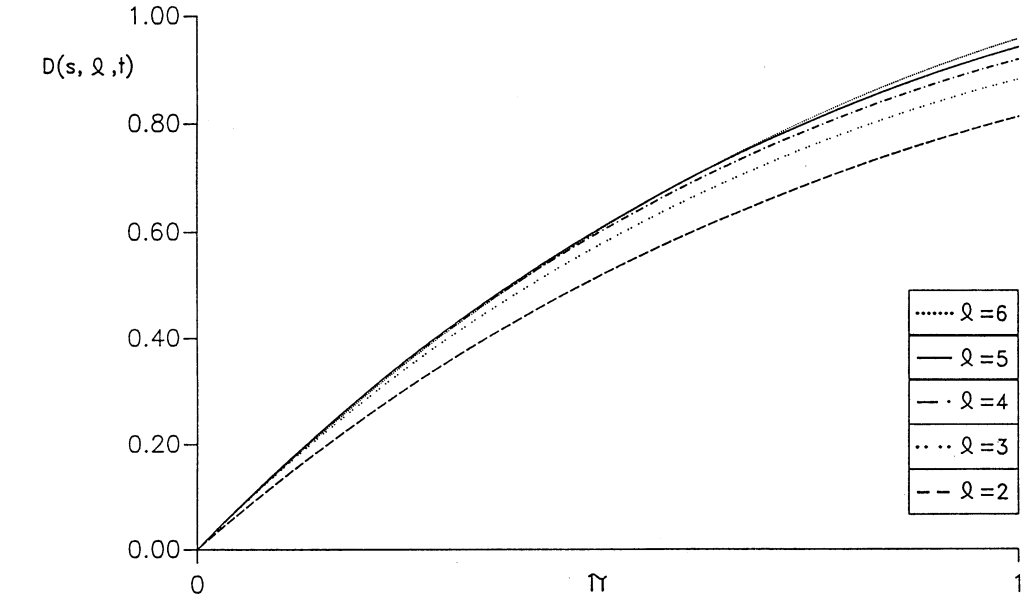


Fig. 2. Detection probabilities for given ℓ and $t = 4$ as a function of π

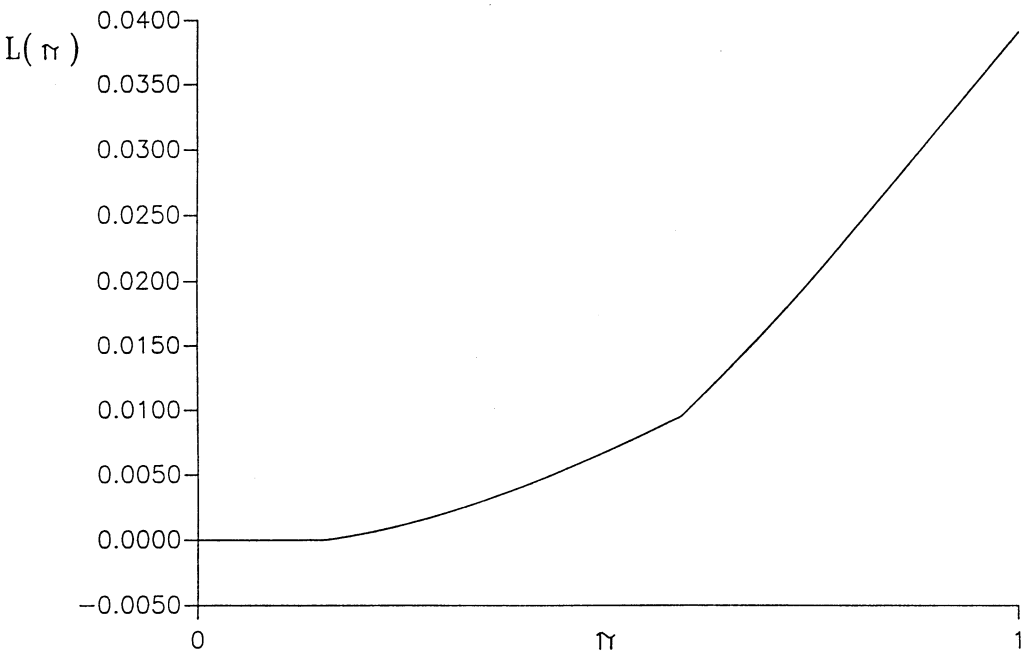


Fig. 3. Loss in detection probability for $\ell = 4$, $t = 4$ as a function of π

Figure 2 illustrates with $p_0 = .657$ the behavior of the detection probability $D(s, \ell, 4)$ as a function of π for $2 \leq \ell \leq 6$. ($p_0 = .657$ when $\delta = 0$ and $\bar{p} = .1$.) It suggests that for low cheating frequency ($\pi < .15$), there is little advantage in moving from the status quo value $\ell = 4$. For more consistent cheaters ($\pi > .15$), the advantage is more pronounced. When p_0 was varied over the range $.4 \leq p_0 \leq .8$, a similar pattern emerged.

Of particular interest is the status quo value of $\ell = 4$. Figure 3 shows the loss in detection probability for $\ell = 4$ as a function of π with $p_0 = .657$, i.e., $L(\pi) = \max_{\ell} D(s, \ell, 4) - D(s, 4, 4)$, where as before s is chosen to satisfy the cost constraint. The greatest loss in detection probability is only about 4%. Because of the advantages of maintaining the current procedure, the reinterview design chosen was $t = 4$ and $\ell = 4$. s was then chosen to satisfy the cost constraint, which resulted in a supplementary sample of $s = I/12$.

6. A Concluding Note

The investigation to date has led us to conclude that no change in the basic design of the reinterview sample (except for the introduction of a supplementary sample) is warranted at this time. The increase in detection probability which might be gained from changing ℓ or t was shown by our model to be too small to justify the added expense and disruption of data collection which would follow a change in reinterview design parameters. Data on cheating cases are still being collected, however, and better estimates of parameters of the cheating model will eventually be available. When this data are available, the optimal design can be reassessed.

In addition to reassessing the optimal choice of s, ℓ, and t for the current sampling scheme, the next redesign of the Census Bureau reinterview program should investigate alternative sampling schemes, especially unequal probability sampling designs. For example, the data in Section 2 indicate that even higher detection probabilities may be realized if the less experienced interviewers or those working in urban areas, or both, were sampled at a higher rate than other interviewers. However, two issues to consider here are: (a) the effect on respondent cooperation for those interviewers sampled more frequently and (b) the effect of the design as a deterrent for those interviewers sampled less frequently. Our current model is insufficient for evaluating these effects.

Appendix 1

Let $P_t(\mathbf{p}, \pi)$ denote the conditional probability that a $(\mathbf{p}, \pi)$ cheater is detected given that he/she is selected for verification in an interviewing period using a verification design having parameters (s, ℓ, t). Define $\boldsymbol{\ell} = [\ell_0, \ldots, \ell_m]$ to be the sample analog to $\mathbf{b}$; i.e., $\ell_h (h = 1, \ldots, m)$ is the number of sample USU's in an interviewer's assignment containing h falsified units.

We assume that $\mathbf{b}$ is distributed as a multinomial random vector with parameters n and $\mathbf{p}$ and denote this distribution by $P(\mathbf{b})$. Thus, since the ℓ USU's are sampled using simple random sampling without replacement, $P(\boldsymbol{\ell}|\mathbf{b})$, the conditional distribution of $\boldsymbol{\ell}$ given $\mathbf{b}$, is the multivariate hypergeometric distribution.

Let A_t denote the event "no falsified interview units detected after inspecting t interview units in each of the ℓ sample USU's". Define the summation Σ' as the sum over all possible $\mathbf{b}$ such that $b_h \geq 0$ and $\Sigma b_h = n$. Likewise define Σ'' to be the sum over all possible $\boldsymbol{\ell}$ such that $0 \leq \ell_h \leq b_h$ and $\Sigma \ell_h = \ell$. Let B be the event "the interviewer is susceptible to cheating" and let B^c denote the complement of B. Now, $P(A_t) = (1 - \pi) + \pi P(A_t|B)$ since $P(A_t|B^c) = 1$. Further,

$$P(A_t|B) = \sum' \sum'' P(A_t|B,\ \mathbf{b},\ \boldsymbol{\ell})\ P\ (\boldsymbol{\ell}|\mathbf{b}) P(\mathbf{b}) \tag{A.1}$$

where by the above assumptions,

$$P(\mathbf{b}) = \binom{n}{b_0,\ldots,b_m} p_0^{\,b_0} \ldots p_m^{\,b_m}, \tag{A.2}$$

$$P(\boldsymbol{\ell} \mid \mathbf{b}) = \frac{\binom{b_0}{\ell_0} \cdots \binom{b_m}{\ell_m}}{\binom{n}{\ell}}, \tag{A.3}$$

and

$$P(A_t|B,\ \mathbf{b},\ \boldsymbol{\ell}) = \prod_{h=0}^{m} \left[\frac{\binom{(m-h)}{t}}{\binom{m}{t}} \right]^{\ell_h}. \tag{A.4}$$

Substituting these into (A.1) yields

$$P(A_t|B) = \sum' \sum'' \prod_{h=0}^{m} \left[\frac{\binom{(m-h)}{t}}{\binom{m}{t}} \right]^{\ell_h} \frac{\ell!}{\ell_0!\ldots\ell_m!} \frac{(n-\ell)!}{(b_0-\ell_0)!\ldots(b_m-\ell_m)!} p_0^{\,b_0} \ldots p_m^{\,b_m}. \tag{A.5}$$

Letting $a_h = b_h - \ell_h$, $h = 1, \ldots, m$, we see that the term involving the b_h under Σ' is

$$\sum' \frac{(n-\ell)!}{a_0!\ldots a_m!} p_0^{\,(a_0+\ell_0)} \ldots p_m^{\,(a_m+\ell_m)} = p_0^{\,\ell_0} \ldots p_m^{\,\ell_m}. \tag{A.6}$$

Thus,

$$\begin{aligned} P(A_t|B) &= \sum'' \prod_{h=0}^{m} \left[\frac{\binom{(m-h)}{t}}{\binom{m}{t}} \right]^{\ell_h} \frac{\ell!}{\ell_0!\ldots\ell_m!} p_0^{\,\ell_0} \ldots p_m^{\,\ell_m} \\ &= \sum'' \frac{\ell!}{\ell_0!\ldots\ell_m!} \prod_{h=0}^{m} \left[\frac{\binom{(m-h)}{t}}{\binom{m}{t}} p_h \right]^{\ell_h} \\ &= \sum'' \frac{\ell!}{\ell_0!\ldots\ell_m!} (p_h^*)^{\ell_h} \end{aligned} \tag{A.7}$$

by the Multinomial Theorem, where $p_h^* = p_h \binom{(m-h)}{t} \Big/ \binom{m}{t}$. Therefore,

$$\begin{aligned} P_t(\mathbf{p},\ \pi) &= 1 - \{(1-\pi) + \pi P(A_t|B)\} \\ &= 1 - \left\{ (1-\pi) + \pi \left[\binom{m}{t}^{-1} \sum_h \binom{(m-h)}{t} p_h \right]^{\ell} \right\}. \end{aligned} \tag{A.8}$$

Finally, the probability of not being detected for an entire evaluation period is the probability of not being detected in the predesignated sample times the probability of not being detected in any supplementary sample. The former probability is simply $1 - P_t(\mathbf{p},\ \pi)$. The latter probability is

$$\sum_{k=0}^{f-1} \Pr(\text{interviewer is selected } k \text{ times and not detected each time})$$

$$= \sum_{k=0}^{f-1} \binom{f-1}{k} \left(\frac{fs}{I(f\text{-}1)}\right)^k \left(1 - \frac{fs}{I(f\text{-}1)}\right)^{f-k-1} [1 - P_t(\mathbf{p}, \pi)]^k$$

$$= \left[1 - \frac{fs}{I(f-1)} P_t(\mathbf{p}, \pi)\right]^{f-1}.$$

Thus,

$$D(s, \ell, t) = 1 - \Pr(\text{not being detected for an entire evaluation period})$$

$$= 1 - [1 - P_t(\mathbf{p}, \pi)]\left[1 - \frac{fs}{I(f-1)} P_t(\mathbf{p}, \pi)\right]^{f-1}.$$

Appendix 2

Theorem 1. If t and $K_1(s, \ell, t)$, defined in (3.7), are held fixed and $\pi = 1$, then $D(s, \ell, t)$ is maximized by either $s = 0$ or $s = \frac{f-1}{f} I$, where its values are equal.

Proof: Let us hold $K_1(s, \ell, t)$ fixed at C_F, i.e.,

$$\gamma(I + fs)\ell t = C_F \tag{B.1}$$

for some constant γ. Then it can be easily shown by substitution into (3.3) and (3.4) that $D(0, \ell, t) = D\left(\frac{f-1}{f} I, \ell, t\right) = 1 - \eta_t^{C_F/\gamma It}$.

Next, to establish that the maximum value of $D(s, \ell, t)$ over the range $0 \le s \le \frac{f-1}{f} I$ is $1 - \eta_t^{C_F/\gamma It}$, we will show that

$$\frac{1 - D(s, \ell, t)}{\eta_t^{C_F/\gamma It}} \ge 1 \tag{B.2}$$

for any fixed t, where $\ell = C_F/\gamma t(I + fs)$ from (A.1). To simplify notation, we write $r = s/[(f-1)/f]I$, so that $0 \le r \le 1$, and $g = C_F/\gamma It[1 + r(f-1)]$. Then the left hand side of (A.2) may be written

$$\left[\frac{1 - r(1 - \eta_t^g)}{\eta_t^{gr}}\right]^{f-1} = \left[(1 - r)\eta_t^{-gr} + r\eta_t^{g(1-r)}\right]^{f-1}. \tag{B.3}$$

Now consider the random variable X having probability function

$$h(x) = \begin{cases} r & \text{for } x = 1 - r \\ 1 - r & \text{for } x = -r \end{cases}$$

and the function $f(x) = \eta_t^{gx}$. Since $f(x)$ is convex, we know by Jensen's inequality that $E(f(X)) \ge f(E(X))$. Since $E(f(x))$ is given by the expression inside the brackets in (A.3) and since $E(X) = 0$, (A.2) is established.

Theorem 2. If t and $K_2(s, \ell, t)$, defined in (3.8), are held fixed and $\pi = 1$, then $D(0, \ell, t) \geq D(\frac{f-1}{f} I, \ell, t)$.

Proof: Define $h_0(\ell) = K_2(0, \ell, t) = I(C_2 \ell t + C_3 \sqrt{\ell} + C_4)$ (B.4)

and $h_1(\ell) = K_2\left(\frac{f-1}{f} I, \ell, t\right) = f\, I\, (C_2 \ell t + C_3 \sqrt{\ell} + C_4)$. (B.5)

Now let ℓ_0 be such that $h_0(\ell_0) = C_F$ and ℓ_1 be such that $h_1(\ell_1) = C_F$. Then

$$\begin{aligned} h_1(\ell_1) = C_F &= fI(C_2 \ell_1 t + C_3 \sqrt{\ell_1} + C_4) \\ &= I(fC_2 \ell_1 t + fC_3 \sqrt{\ell_1} + fC_4) \\ &> I(C_2 f \ell_1 t + C_3 \sqrt{f\ell_1} + C_4) \text{ since } f > 1 \text{ and } C_3,\ C_4 > 0 \\ &= h_0(f\ell_1). \end{aligned}$$

Since $h_0(f\ell_1) < C_F$, $h_0(\ell_0) = C_F$, and $h_0'(\ell) > 0$, we know $f\ell_1 < \ell_0$. Thus by substitution into (3.3) and (3.4) we have

$$D(0, \ell_0, t) - D\left(\frac{f-1}{f} I, \ell_1, t\right) = \eta_t^{f\ell_1} - \eta_t^{\ell_0} > 0,$$

and the theorem is established.

7. References

Bennett, A. (1948a): Survey on Problems of Interviewer Cheating: Observations on the So-Called Cheater Problem Among Field Interviewers. International Journal of Opinion and Attitude Research, 2, pp. 89–96.

Bennett, A. (1948b): Toward a Solution of the "Cheater Problem" Among Part-Time Research Investigators. Journal of Marketing, 2, pp. 470–474.

Biemer, P., Judkins, D., Schreiner, I., and Stokes, S. L. (1982): Reinterview Redesign Report 1: Optimal Sampling Strategy for Interviewer Control. Bureau of the Census Technical Report, U. S. Bureau of the Census, Washington, D. C.

Boyd, H. and Westfall, R. (1955): Interviewers as a Source of Error in Surveys. Journal of Marketing, 19, pp. 311–324.

Crespi, L. P. (1945): The Cheater Problem in Polling. Public Opinion Quarterly, Winter, pp. 431–445.

Evans, F. B. (1961): On Interviewer Cheating. Public Opinion Quarterly, 25, pp. 126–127.

Pielou, E. C. (1969): An Introduction to Mathematical Ecology. John Wiley and Sons, Inc., New York.

Sheatsley, P. B. (1951): An Analysis of Interviewer Characteristics and Their Relationship to Performance, Part II. International Journal of Opinion and Attitude Research, 5, pp. 79–94.

U. S. Bureau of the Census (1986): Memorandum for General Distribution from I. Schreiner and K. Guerra, Subject "Falsification Study" (September 1982 through August, 1985).

Received March 1988
Revised March 1989

Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 41–55

Early Survey Models and Their Use in Survey Quality Work

Gösta Forsman[1]

Abstract: There have been great advances in sampling models over the past 60 years. As these models have been developed, so has an awareness of the problem of nonsampling errors in surveys. Two lines have emerged in this work, namely (i) the development of theory and methods for handling specific sources of nonsampling errors, and (ii) the development of a comprehensive theory of an integrated treatment of survey errors. The latter line is characterized by the use of survey models. This paper deals with the early research on survey models up to the early 1970s, and looks at the application of these models in survey quality work.

Key words: Nonsampling errors; survey models; survey quality.

1. Introduction

At a statistical agency, survey quality work includes a variety of procedures such as evaluation studies, preventive control, and production control. One fundamental part of this work is the measurement of survey errors. Measurement studies provide information about quality that is useful for both the producer of the data and the user. The survey methodologist needs data on survey quality to improve methods and to allocate resources more effectively. The user of the statistics needs quality data to determine whether the survey estimates are reliable enough to meet his/her needs.

The early development of survey theory focused on the measurement and control of specific error sources. An important example is the very successful research on sampling errors. Since the 1930s there has been an increasing awareness of the problems of nonsampling er-

[1]Senior statistician, Statistical Research Unit, Statistics Sweden, S-115 81 Stockholm, Sweden. This paper is adapted from Forsman (1987).

Acknowledgements: I would like to thank the following persons who kindly supplied information. C. Andersson, B. A. Bailar, S. Berg, P. P. Biemer, T. Dalenius, J. C. Deville, L. Fabbris, I. Fellegi, G. B. Gray, M. Hansen, W. Keller, L. Kish, G. Koch, J. Koop, J. T. Lessler, I. Lyberg, L. Lyberg, R. Platek, J. N. K. Rao, M. Ribe, H. Strecker, P. V. Sukhatme, G. Theodore, A. Sunter, V. Verma, and S. S. Zarkovich. The editor and two anonymous referees provided many helpful suggestions. I am, however, responsible for any defects that might be found in the paper.

rors. India and the United States led the early development in this field. Forerunners were the Indian Statistical Institute, led by Mahalanobis, the Indian Council of Agricultural Research, where P. V. Sukhatme worked, and the U. S. Bureau of the Census, where Hansen, Hurwitz, Tepping, Madow, and others were pioneers. Contributions from the United Kingdom were also important. At the Rothamstead Experimental Station, Fisher, Yates, Cochran, and others did research on statistical experiments in the 1920s and 1930s. This work had a strong influence on the development of survey theory.

Studies of specific error sources have continued to be an important part of survey quality work. In the 1940s a parallel development emerged that aimed at an integrated control of all sources of errors and thus of the total error. In this research, what is called mixed error models were developed; later, the term survey models has been widely used.

We can distinguish three fields of application for survey models:

1. As already indicated, a survey model allows an integrated treatment of various error sources. Thus using the model, the total error can be estimated. Note that the total error given the model is the error resulting from the error sources that the model takes into account. The "real" total error of a survey estimate may be affected also by other sources of error.

2. Survey models can be used to estimate the relative impact of different error sources on the total error. For recurrent surveys, this allows a reallocation (if necessary) of resources to effectively control the error sources.

3. Survey models might also be applied to a specific source of error to study the magnitude of its components. For example, if applied to the response error, we can estimate the total response error and its components, the response bias and the response variance. This also can lead to an improved allocation of the resources among survey operations.

We will review the early development of survey models up to the early 1970s and discuss their use in the subsequent work on survey quality. It is mainly these models and modified versions of them that have been used in survey practice so far. The presentation is restricted to models that include estimation procedures for the error components. The theoretical development is reviewed in Section 2. Section 3 contains examples from survey practice and in Section 4 we discuss the application, or lack of application of survey models.

2. Survey Models

2.1. Models for variable measurement errors

The work to develop survey models (led by Indian and American statisticians) concentrated on sampling variance and measurement variability. Two important sources of measurement variability were identified early:

a) the error that depends on the tendency of the interviewers (or enumerators or observers, depending on the data collection mode) to affect the respondent's answers, and

b) the error that emerges from the fact that the answers to a question can be different if the same respondent is asked the same question on different occasions.

There has not yet emerged a common nomenclature for these sources of error. In this section, I refer to them as interviewer error and respondent error, respectively. It should be noted that the respondent error includes some of the effect of the interviewer error if different interviewers ask the same questions on the two occasions.

In the United States, Rice (1929) showed that the interviewers' own attitudes affected the respondents and could lead to a response error. There was an urgent need to measure this type of error at the U.S. Census Bureau. The data collection in the Decennial Censuses of Population and Housing was conducted by thousands of temporarily employed interviewers whose skills could vary considerably. This type of error was also well-known in India, e.g., in the crop surveys where the observers might classify the same field very differently.

The modeling of the interviewer error was done somewhat differently in different agencies. Central to the U.S. Census Bureau's model was that each interviewer generated clusters of responses. Then, by allocating a random subsample to each interviewer, the error could be measured by a cluster sampling (variance) formula. The situation here differs, however, from that in cluster sampling. In cluster sampling, the correlation seen in the data reflects the correlation that exists in the population. In the U.S. Census Bureau's model, on the other hand, the correlation is a result of the observation and data collection process. This error component was called the correlated response variance. In India and in some other agencies in the United States, this error was regarded as a bias due to the interviewer (or the observer). The interviewers were considered a simple random sample from a population of interviewers. The variance among the biases associated with these interviewers in the population was often called the interviewer variance. This variance component could then be estimated from the sample of interviewers. The interviewer variance is often regarded as identical to the correlated component of the response variance according to the Census Bureau model. This is only approximately true, however, since, theoretically, the correlated component of the response variance can take on negative values.

The modeling of the respondent error was also done differently at different agencies. At the U.S. Census Bureau the random nature of this error was discussed early (see Palmer (1943), and Deming (1944)). In the model developed at the Census Bureau it was assumed that an answer to an interview question is generated by a random process. As a consequence – even a response from a given respondent to a given interviewer has a probability distribution. A similar situation is assumed in Sukhatme (1954) and Sukhatme and Seth (1952) and probably also in Mahalanobis (1946), although Mahalanobis does not explicitly describe a survey model as we have defined it here. Another way of modeling the respondent error is to assume that only one answer is possible for each respondent-interviewer-question combination. A given respondent could, however, provide different answers to the same question to different interviewers. The stochastic element in a survey model with this assumption is entirely due to the sampling processes and the allocation of respondents to interviewers. Both interviewers and respondents are usually regarded as sampled from large populations. One can interpret the survey model described by Stock and Hochstim (1951) as based on this deterministic approach. The same goes for the later model by Murthy (1967) and the conceptual discussion in Zarkovich (1966).

The models for the total survey error, which considered the sampling error and the two variable measurement error types described above, were usually formulated according to two basic ideas. The Census Bureau used a mean square error decomposition approach founded in sampling theory, while other agencies used a linear model approach founded in the analysis of variance (ANOVA) technique.

2.1.1. The mean square error decomposition approach

The Census Bureau model assumes a set of general conditions under which the survey is conducted. The survey is regarded as one trial

from among a large set of conceived repetitions of the survey under the same general conditions. This means that a measurement derived from the survey has a well-defined, but unknown, probability distribution. The model postulates the existence of a true value, x, for each sampling unit. We denote the measurement for the ith element at the tth trial by y_{it}. Now, the conditional expected value of y_{it} over all possible samples that include the ith element and all possible trials that have resulted in such a sample, is

$$E(y_{it}|i) = Y_i. \tag{2.1}$$

The difference between the observation on the ith unit in a particular survey and the conditional expected value of that unit is

$$d_{it} = y_{it} - Y_i.$$

d_{it} is called the response deviation.

Assume now, that in a specific trial, t, the population mean, $\bar{X}$, is to be estimated by $\bar{y}_t$, the sample mean from a simple random sample of n units. Then the total error $\bar{y}_t - \bar{X}$ is measured by $\mathrm{MSE}(\bar{y}_t)$, which can be decomposed as:

$$\mathrm{MSE}(\bar{y}_t) = \sigma_S^2/n + \sigma_R^2[1 + (n-1)\varrho]/n + 2(n-1)\sigma_{RS}/n + B^2. \tag{2.2}$$

In (2.2), the first term is the sampling variance of $\bar{y}_t$, defined as the variance among the Y_i-values in the population, divided by n. The second term is the response variance, defined as the variance of $\bar{d}_t$, the average of the response deviations for the sample. This term can be further decomposed into the simple response variance, σ_R^2/n, (which is the error component corresponding to the respondent error) and the correlated response variance, $\varrho(n-1)\ \sigma_R^2/n$. Here, ϱ is the intraclass correlation coefficient among the response deviations for a trial (survey), defined as

$$\varrho = E(d_{it}, d_{i't})/\sigma_R^2, \quad i \neq i'.$$

It is important to recall that the sampling variance measures variations caused by the sampling process, while the response variance measures variations assumed to characterize the measurement operation. The third term in (2.2) is the covariance of the response and sampling deviations, which is normally regarded as very small – it is zero for a complete census. The fourth term, finally, is the squared bias.

An important feature of the model is its broad applicability. It may be applied to any sequence of survey operations, i.e., either the full sequence or a subset of operations (for instance, interviewing and coding). Applied to the full sequence, the response variance reflects contributions from all operations such as interviewing, coding, editing, and so forth. Applied to coding alone, the response variance reflects only coding and the response variance becomes a coding variance. Analogously to (2.2), coding gives a contribution to the MSE of the form

$$\sigma_C^2[1 + (n-1)\varrho_C]/n + B_C^2. \tag{2.3}$$

For surveys with interviewers, the correlated response variance may be especially large. It is then important to note that this component does not decrease when the number of sampled units within an interviewer's assignment increases. Hence a relatively low value of ϱ can have a considerable effect on the total response variance and also on the total MSE. This is readily seen if the correlated response variance is assumed to depend entirely on the interviewers (i.e., on the "interviewer error" described in Section 2.1). The response variance is then given by

$$\sigma_R^2[1 + \varrho(m-1)]/n, \tag{2.4}$$

where m is the (average) number of respondents assigned to an interviewer.

The Census Bureau survey model was first presented in Hansen, Hurwitz, Marks, and Mauldin (1951). In this paper, Hansen et al.

assumed that the correlated response variance depends entirely on the interviewers. They showed that the correlated component of the response variance could be estimated by means of interpenetrating subsamples. They also showed that the common textbook estimator of the sampling variance of $\bar{y}_t$ actually estimated the sum of the sampling and simple response variances. During the 1950s, the model was further elaborated on and eventually presented in the two widely recognized papers by Hansen, Hurwitz, and Bershad (1961) and Hansen, Hurwitz, and Pritzker (1964). In these papers, the correlated component of the response variance was defined as dependent not only on the interviewers, but on all field personnel. The correlation between answers to different interviewers was permitted to be nonzero, reflecting the possible correlations arising from supervisors, coders, editors, keyers, etc. The above assumptions apply to (2.2).

In Hansen et al. (1964), an estimator of the simple response variance for 0,1-variables was presented. This estimator was derived under the assumption that independent repeated measurements of the sampled units were conducted. In this case the original survey and its replication (the reinterview) were assumed to be two independent randomly selected trials. Now the gross difference rate,

$$g = \sum_{i=1}^{n} (y_{i1} - y_{i2})^2/n, \tag{2.5}$$

divided by two, can be shown to be an unbiased estimate of σ_R^2/n. Hansen et al. (1964) defined an index of inconsistency as the ratio of the simple response variance to the total variance of individual responses, σ_y^2, that is

$$I = \sigma_R^2/\sigma_y^2. \tag{2.6}$$

For a Bernoulli random variable with parameter P, the total variance σ_y^2 is $P(1-P)$. An estimator of the numerator is then $g/2$. The denominator may be estimated by $\bar{y}_t(1-\bar{y}_t)$, $t=1, 2$. Here $\bar{y}_t$ is the proportion of sample units that belongs to the category of study in trial t data from either trial 1, or trial 2, or from both trials may be used). Obviously, the index takes values between 0 and 1. Low values of the index indicate that the measurement process is under control.

The two procedures for estimating error components mentioned above, interpenetration and repeated measurements, are the basic methods available in the estimation process. Fellegi (1964) demonstrated how the two procedures could be combined. In his notation, the assignment of the jth interviewer is $S_{j(1)}$, $S_{j(2)}$, $j=1, \ldots, k$, where $S_{j(1)}$ and $S_{j(2)}$ are randomly allocated assignments for the jth interviewer in the original and reinterview surveys, respectively. $S_{j(1)}$ and $S_{j(2)}$ are not the same for a given interviewer. The model is similar to the Hansen, Hurwitz, and Pritzker (1964) model but differs in that the conditional expected values of a measured value y_{ijt} given a respondent, i, and an interviewer, j, over the trials need not be the same for the original survey and the reinterview. Fellegi's data collection design accomodates the definition of several types of correlation among the response deviations.

Bailar and Dalenius (1969) demonstrate the potential usefulness of the procedures interpenetration and repeated measurements. The procedures are reviewed (also in combination) in several basic study schemes aiming at estimating different variance components in the Census Bureau model. The study schemes are classified according to repetition and interpenetration in two dimensions, called the sample dimension and the trial dimension. Repetition in both dimensions is characterized by the use of the same sample and the same field personnel (e.g., interviewers and coders) in a replicated study, repetition in the sample dimension combined with interpenetration in the trial dimension implies the use of the same sample and different field workers in a replicated study, etc. Some of the study schemes are even more sophisticated than Fellegi's, but these schemes are also more difficult to implement.

Sometimes, when interpenetration or repetition cannot be applied in a survey, data from other surveys are used instead. For example, data from a match between a labor force survey and a census with labor force items may sometimes be analyzed as if they were reinterview data. Moreover, data from two independently conducted surveys with similar questions and data collection procedures on the same population may be treated as if interpenetration had in fact been applied (see Tepping and Boland (1972)).

Repeated measurements can also be used to estimate the bias component. The measurements must, however, be carried out with a preferred procedure that can be assumed to provide data close to the true values.

Murthy (1967) and Des Raj (1968) present variance decomposition models. They both arrive at expressions of the variance of the sample mean that are similar to the Census Bureau model decomposition, although their components have different definitions. Murthy conceives of the survey as having two steps of randomization:

i. a sample of population elements, s, and

ii. a sample of survey personnel, r.

He defines y_{ij} as the value obtained by the jth interviewer for the ith element. Since Murthy assumes the deterministic response model described in 2.1, this value is not a random variable. He gives the following expression for the variance of the sample mean, $\bar{y}$:

$$V(\bar{y}) = \sigma_s^2/n + \sigma_d^2[1 + (m-1)\,\varrho]/n,$$

where the terms are called sampling variance, simple or uncorrelated response variance, and correlated response variance, respectively. The names are the same as the names of the components of the Census Bureau model, but they are not the same components; the response deviations are defined differently.

Contrary to Murthy, Des Raj defines the observed value y_{ij} as a random variable. In his design the interviewers are allocated randomly to primary sampling units which have been selected with probabilities proportional to size. Thus, Des Raj is the first to present a survey model based on a PPS sampling design.

2.1.2. The linear model approach

Linear survey models may also be constructed in many different ways. The models emerged from the analysis of variance theory. In the late 1930s and early 1940s, the Indian Statistical Institute was the first to use ANOVA-type models in survey practice. Under the leadership of P.C. Mahalanobis sampling designs for crop surveys with embedded experiments based on interpenetrating subsamples were developed. The purpose of these experiments was to control the individual investigator bias that had been encountered in surveys where different investigators had unknowingly been allotted the same fields.

One of the first examples in the literature of measuring the overall error by means of linear survey models was provided by Stock and Hochstim (1951). The authors seem to assume the deterministic response model as mentioned above, i.e., only one answer is possible for each respondent-interviewer-question combination. In this application, the deterministic model presupposes a population of N respondents and a population of K interviewers. Then, assuming that each interviewer, j, interviews each respondent, i, in the population, the data generating process may be modeled in the following way:

$$y_{ij} = \bar{y}.. + I_j + e_{ij},\ i=1, .., N,\ j=1, .., K,$$

where

$\bar{y}..$ is the mean of all $N \times K$ observations,

I_j is the individual interviewer bias for interviewer j, i.e., the difference between $\bar{y}..$ and

the mean of the N observations, $\bar{y}_{.j}$, of interviewer j, and,

e_{ij} is the deviation between the observed value and $\bar{y}.. + I_j$ when interviewer j interviews respondent i.

It is assumed that no correlation is present between I and e. The survey design is such that an interpenetrated subsample, drawn by simple random sampling, is randomly allotted to each of k interviewers. The interviewers are sampled by simple random sampling from a population of K interviewers.

Now, if assuming that the sampling fractions are small, the variance of the sample mean, $\bar{y}$, is approximately:

$$V(\bar{y}) = \sigma_I^2/k + \sigma_e^2/n,$$

where n is the total sample size.

σ_I^2 is the variance among the I_js, defined by $\Sigma_j^K (\bar{y}_{.j} - \bar{y}..)^2/(K-1)$. This is the interviewer variance mentioned in Section 2.1, and it is present in most ANOVA-type survey models (although it may be slightly differently defined). σ_e^2 is the variance between respondents within interviewers, averaged over all interviewers. σ_e^2/n is the sampling variance and would be the total variance under the model if there were no interviewer effects. Note that the model does not take into account the existence of true values.

Sukhatme (1954) presents a linear survey model different from the Stock and Hochstim model in that (i) the existence of true values is assumed, and (ii) one of the error components (and thus y_{ij}) is regarded a random variable. Like Stock and Hochstim, Sukhatme assumes finite populations of N respondents and K interviewers. He lets

$$y_{ij} = x_i + \alpha_j + \varepsilon_{ij},$$

where x_i is the true value. α_j is defined as "the bias of the jth enumerator in repeated observations on all units." α_j is very close – if not identical – to I_j in the Stock and Hochstim model. ε_{ij} is the random deviation of the reported value from $x_i + \alpha_j$. It is assumed that the ε_{ij}s are independently distributed with mean 0 and variance σ_ε^2 for all $i = 1, \ldots, N$ and $j = 1, \ldots, K$. Assuming the same sampling and measurement design as in the above description of the Stock and Hochstim model, Sukhatme derives the following expression for $V(\bar{y})$:

$$V(\bar{y}) = \sigma_x^2 (1/n - 1/N) + \sigma_\alpha^2 (1/k - 1/K) + \sigma_\varepsilon^2/n,$$

or, if N and K are large:

$$V(\bar{y}) \approx (\sigma_x^2 + \sigma_\varepsilon^2)/n + \sigma_\alpha^2/k$$

σ_x^2 is the variance among the true values in the population. It cannot be estimated separately from σ_ε^2 with Sukhatme's design. Sukhatme shows that the mean square between observations within enumerators is an estimator of the sum of σ_x^2 and σ_ε^2. This result is similar to the finding that the common sampling variance estimator estimates the sum of the sampling and simple response variances in the Census Bureau model. However, the components are differently defined in the two models. σ_α^2 is, analogously to σ_I^2 above, defined as $\Sigma_j^K (\alpha_j - \bar{\alpha})^2/(K-1)$, where $\bar{\alpha} = \Sigma_j^K \alpha_j/K$.

Sukhatme also derives the correlation, ϱ', between responses obtained by the same interviewer. If the finite population correction is small, the variance of a single observation is approximately $\sigma_y^2 = \sigma_x^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2$. An approximate expression for ϱ' is

$$\varrho' = \sigma_\alpha^2/\sigma_y^2 = \sigma_\alpha^2/[\sigma_x^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2],$$

which gives $V(y) = \sigma_y^2 [1 + \varrho'((n/k) - 1)]/n$.

A similar model is presented by Kish (1962), who, however, pools the x and ε components. The reason for this is that (as is mentioned above) the variance components σ_x^2 and σ_ε^2 cannot be estimated separately with Kish's (or Sukhatme's) design since it does not include

repeated measurements. Kish's model can be written

$$y_{ij} = y_{ij}' + \alpha_j,$$

where $y_{ij}' = x_i + \varepsilon_{ij}$. The definitions of the components seem to be the same as in the Sukhatme model. The intraclass correlation, ϱ^*, between responses obtained by the same interviewer is defined as:

$$\varrho^* = s_a^2 / (s_a^2 + s_b^2), \tag{2.7}$$

where s_a^2 and s_b^2 are the sample estimates of the variance between interviewers and within interviewers, respectively.

The Sukhatme (1954) model is a simplified version of the model presented in Sukhatme and Seth (1952). The latter model allows several observations on each unit and an interaction between the interviewer and the respondent.

The above linear models are based on the assumption that the expected values of the measurements do not depend on the sample. This assumption is usually not made in the variance decomposition models. E.g., the Census Bureau model takes the covariance between of the response and sampling deviations into account, see the third term in (2.2).

2.1.3. Other models for measurement variability

There are other approaches to developing survey models for surveys intended to estimate proportions in the presence of classification errors. In these models, the key concept is what is called misclassification probability. Assume, for example, that the true value, x_i, for unit i is 1 if i belongs to some category, say C, and 0 otherwise. Errors that give rise to the misclassification of a unit are considered. The survey procedure is assumed to generate a stochastic variable y_i, such that

$$\varphi = \Pr(y_i = 1 \mid x_i = 0), \text{ and}$$

$$\theta = \Pr(y_i = 0 \mid x_i = 1), \; i = 1, \ldots, N.$$

φ and θ are the misclassification probabilitites.

Casady (1966) derived two such survey models for the analysis of reinterview data in the Health Interview Survey conducted by the U.S. National Center for Health Statistics. Casady defined the "within element response variability" and an index of inconsistency and presented estimators of these parameters using the two models. The models differed in that the misclassification probabilities in the first model were assumed to be constant over different trials while the misclassification probabilities were permitted to vary between trials in the second model.

Swensson (1969) showed that the first Casady model could be regarded as a special case of the Census Bureau model (only the sampling variance and the simple response variance are regarded – the correlated response variance cannot be studied under the given definition of misclassification probabilities).

The misclassification approach was discussed in Cochran (1968), who defined the misclassification probabilities as dependent on the unit. Bailar and Biemer (1984) showed that the misclassification probabilities can be formulated as dependent on both the unit and the operator (i.e., an interviewer, coder, supervisor, etc.). This allows a correlated measurement error component to be estimated using the misclassification probability approach.

2.2. *Survey models for systematic errors*

Systematic errors are normally studied by comparing the survey data to preferred data. Higher quality data are usually obtained from reinterviews or record checks. The method allows an estimation of the bias term as it appears in, e.g., the Census Bureau model (see formula 2.2). This term has been extensively studied within the U.S. census evaluation programs.

In addition to the Census Bureau work on survey models, bias models were developed for specific survey situations. Kish and Lansing

(1954) developed a model for the case where not only the observed values but also preferred values, obtained from a preferred data collection procedure, are available. These preferred values were, however, not regarded as good as the true values. This model was to estimate the error in a study of the market value of houses, a study that was part of the 1950 Survey of Consumer Finances in the United States.

The well-known randomized response model presented by Warner (1965) can be regarded as a survey model since it takes into account different error sources. In Warner's model, response errors with known probabilities are intentionally introduced to eliminate nonresponse and erroneous answers to sensitive questions. This technique makes it possible to construct unbiased maximum likelihood estimators of population means and totals.

The connection between survey models and randomized response models is even more evident in the paper by Abul-Ela, Greenberg, and Horwitz (1967), who extended the Warner model to a trichotomous randomized response model. Contrary to the Warner model, the respondent is assumed to tell the truth with a probability that is allowed to be less than 1 in the Abul-Ela et al. model.

3. Use of the Early Survey Models

Several aspects of survey quality work have been affected by the use of early survey models. These models have been used in regular surveys, in evaluation studies, and in development work. In addition, the early survey models provided a conceptual framework that has completely permeated survey practice. The simple and correlated components of the response variance, the interviewer variance, and the index of inconsistency have become well-known and useful concepts, often theoretically discussed in technical reports even in cases where the models have not been explicitly applied.

In this section we will give concrete examples of work on data quality, guided and inspired by the early models presented above. The examples are confined to the models described in Sections 2.1.1–2.1.2, which are the most frequently used. It should be emphasized that the list is by no means a comprehensive review of the quality work guided and inspired by these models.

For each of the two main approaches for formulating survey models, the MSE decomposition approach and the linear model approach, we saw that different survey models may be developed. However, the two approaches can lead to very similar models which can be used for decomposing the total error into similar components, the same estimators of these components are used and, consequently, also the same data collection designs. The choice between different approaches is then probably more dependent on the way the statistician is used to structuring statistical problems than on other considerations. There is a difference in the use of the approaches. In the quality work performed by governmental agencies, the mean square error decomposition approach is by far the most common. Probably, the large-scale work with these models at the U.S. Bureau of the Census and Statistics Canada has influenced this decision. The linear models are more frequently used in research work on interviewer effects conducted in survey agencies outside the national bureaus.

3.1. Models based on mean square error decomposition

Not very surprisingly, the most extensive quality work based on survey models has been conducted at the U.S. Bureau of the Census. The bulk of the work has been done within two of the Bureau's major projects: the Decennial Census of Population and Housing and the Current Population Survey.

A continuing program of research, evaluation, and experimental studies has been conducted as a part of the censuses and during the intercensal periods. The results of the 1950 census experiments led to important changes in procedures adopted for the 1960 census. In one of these experiments, a set of interviewer-assignment areas was designated. In these areas, the interviewers' assignments were randomly allocated according to the design postulated in the Hansen et al. (1951) model. This experiment dealt with the variance between and within interviewers. The intraclass correlation of response errors within interviewers was also estimated. In U.S. Bureau of the Census (1985), these intraclass correlations are reported for items such as race, age, educational attainment, income, etc. Among the items having the largest ϱ's were the not-reported-categories indicating the influence of enumerators on item nonresponse rates.

For items that are typically difficult to measure (i.e., occupation, education, and income) the correlation was often around .03 (see also Hansen and Tepping, 1969, p. 11). This seems small, but when the average size of an interviewer's assignment is about 700, the factor $[1 + \varrho\,(m-1)]$ in (2.4) becomes larger than 20, leading to a substantial contribution to the total variance even for a moderate σ_R^2. These and similar findings showed that the variability in the complete census results was as large as if only a 25 % sample had been taken (in the absence of interviewer effects). This was true even for areas with populations smaller than 5 000 people. These findings along with studies of the bias and experimental studies of self-enumeration, etc., led to the following procedural changes for the most difficult items to measure in the 1960 census:

i. The data collection was based on a 25 % sample.

ii. A self-enumeration procedure was introduced for this sample.

The interviewers were, however, still engaged in the data collection for the 1960 census. Interviewers delivered the questionnaires to the households and completed them for those households that did not mail in a completed form or whose questionnaires were inconsistent. This led to an interviewer influence on the variance in the 1960 census too. It was much smaller than in the 1950 census, but, however, still important for a number of items.

In the 1970 and 1980 censuses, changes were made in the census-taking procedures in that the questionnaires were delivered by mail to most of the population (95 % in the 1980 census). The enumerators still had an important role in the follow-up procedures, and enumerator variance studies were also made in the evaluation programs of these censuses.

Within all content evaluation programs of the censuses from 1950 to 1980, large-scale reenumeration studies were conducted to obtain estimates of response variance and bias. The reenumerations were conducted as reinterviews or as a record match to the Current Population Survey.

The Census Bureau model was also applied to the coding process, as described above in formula (2.3). Jabine and Tepping (1973) presented estimates on the simple and correlated coding variance components (presented as relvariances) for 1960 census data. These were related to sampling and total response relvariances, as well as to response and coder bias (the latter was based on 1970 census data).

In the Current Population Survey (CPS) a continuing reinterview program has been conducted since the early 1950s. These studies are primarily designed to control the field procedures, rather than measuring the simple response variance according to the Census Bureau model. Nevertheless, reinterview data are continually used to derive the index of inconsistency for various items. According to U.S. Bureau of the Census (1978), this measure has an important role in CPS quality work: "The index

is used primarily to monitor the measurement procedures over time. Substantial changes in the indexes that persist for several months result in review of field procedures to determine and remedy the cause."

Experiments aiming at measuring the correlated components of the response variance are not conducted in the CPS. Tepping and Boland (1972) report, however, from a study where data from the Monthly Labor Survey, carried out during six months in 1966 concurrently with the CPS, provided estimates independent of the regular CPS estimates for several items. The two estimates could then be used for estimating the correlated response variance component. In this paper, Tepping and Boland present estimates of the ratio of the correlated response variance to the sum of the sampling variance and the simple response variance, i.e., in terms of Section 2.1:

$$(m-1)\,\varrho\,\sigma_R^2/(\sigma_R^2+\sigma_S^2).$$

The estimated ratios range between 0.5 and 1.0.

In Canada, the Fellegi model was applied in an experimental pilot study preceeding the 1961 Canadian Census of Population. The results were similar to those found in the U.S. census in that the correlated response variance, derived as the mean of the correlated response variances in the two surveys, was "several times as large as the simple response variance for all except the basic population counts, such as the number of males, sons, married persons, persons of certain age, etc." (Fellegi (1964). Fellegi concluded that, for most characteristics, "considerable gains in the total response variance may be made by reducing the size of the enumerators' assignments". Fellegi argued that the Canadian Census should use a self-enumeration procedure. To determine if such a procedure would increase the simple response variance, he compared the index of inconsistency for a self-enumeration survey with the index of inconsistency for an interview survey. He used items from the 1960 U.S. census (self-enumeration) that corresponded to items in his pilot study for the 1961 Canadian census (interviews). Fellegi found that the values of the simple response variance were rather similar despite the different procedures. As a result of these findings, the 1971 Population Census of Canada was substantially modified. Self-enumeration was introduced along with a sample based collection of most census questions. Later, Krotki and Hill (1978) compared the Fellegi estimates of the correlated response variance with the corresponding estimates from the 1971 and the 1976 Canadian censuses. They found that for almost all characteristics examined, the magnitude of the estimates were considerably reduced.

In Spain, an evaluation program of the General Population Survey (which includes, e.g., labor force items) has been conducted since the early 1970s. The program is based on 3 000 reinterviews each quarter. The purposes of the program are to control the work of the interviewers and to evaluate the general quality of the results. According to Sanchez-Crespo (1973, 1981), the quality evaluation is based on the U.S. Census Bureau model. In the 1981 paper, estimates of the total response variance, the simple response variance, and the correlated response variance are presented for the variable "unemployed" (the study design used for estimating the correlated component is, however, not described). The correlated component was found to give the largest contribution to the total response variance.

In Belgium, a variance decomposition model for surveys with reenumerations, developed by Strecker and Wiegert, was applied in the 1979 Census of Agriculture. The application was limited to one variable, viz., the number of pigs. A study was conducted which provided both replicated data on which estimates of the simple response variance were based, and preferred data, on which bias estimates were

based. The impact of the simple response variance component was considerable, as the following example (from Strecker, Wiegert, and Kafka (1984)) shows. The mean square error was defined as the sum of the simple response variance, the sampling variance, and the squared bias. The relative MSE for the estimate of the mean number of pigs per holding was estimated to 4.61 %. If the relative MSE for this variable had been defined as the sum of the sampling variance and the squared bias only, it would have been 1.92 %. Thus the simple response variance more than doubled the relative MSE.

At Statistics Sweden, the quinquennial Censuses of Population have been evaluated during the last decades. An evaluation based on a survey model is conducted only for labor force items, though. This is based on two sets of data. One data set is created by a match between census labor force items and Labor Force Survey (LFS) data collected during the same time as the census is taken. The LFS data set is then regarded as an independent replication of the census data. The other data set is created by a reconciliation of the census-LFS match and is regarded as preferred data.

Lyberg (1986) reports estimated error components for the items "hours/week at work" and "outside the labor force" for the 1980 census data. In general, the simple response variance is small compared to the squared bias. This fact together with the assumption that the correlated response variance is small (because the census data are collected by mail) has led to the conclusion that bias is the major problem in the Swedish population census. However, neither the impact of the editing personnel on the estimates nor the impact of other items except those mentioned above have been studied.

3.2. *Linear survey models*

In the above mentioned papers by Stock and Hochstim (1951), Sukhatme and Seth (1952), and Kish (1962), examples of studies of interviewer effects are described. In recent years Kish's simple model has been frequently used. The parameter of study in these applications is the intra-class (or intra-interviewer) correlation, ϱ^*, defined by (2.7). The Kish approach is relatively undemanding in terms of experimental design and permits comparisons between studies involving different numbers of interviewers and respondents. We shall review two examples of such studies.

Collins (1980) reports three experiments on interviewer variability conducted by the Social and Community Planning Research (SCPR) in the United Kingdom. The experiments took place in Southampton, North Yorkshire, and Milton Keynes. The questionnaires dealt with the problems faced by the disabled, environmental preferences, and different aspects on living and working, respectively. The estimated interviewer effects (ϱ^*) were generally larger in the Southampton study than in the two other studies. One possible explanation for this is the topic of study. Some categories of question are more prone to interviewer variability than others. Examples of questions prone to interviewer variability would be questions which the interviewer is reluctant to ask and the answer is often imputed from responses given elsewhere in the interview. Such reluctance can be common in a study of disability and its consequences. The results from the North Yorkshire and the Milton Keynes experiments were remarkably similar, despite the fact that the former dealt mainly with attitudinal items and the latter mainly with factual items. This confirmed results from comparisons reported by Kish (1962), who could not find any systematic differences in ϱ^*-values between attitudinal and factual items.

At the Survey Research Center at the Institute for Social Research (ISR), University of Michigan, U.S.A., the Kish model has been frequently applied in measuring interviewer effects, first by Kish in studies of factory workers'

job attitudes. In recent years, Groves and others have applied the Kish model in various telephone surveys. Groves and Magilavy (1986) reviewed nine ISR telephone surveys and the estimates of ϱ^* for 297 survey items. Other interview surveys were also reviewed in which similar models for interviewer effects were applied. The average values of ϱ^* were in eight of the nine ISR surveys under .01, but varied considerably between different statistics. The lowest average of ϱ^*, .0018, was found in the survey with the largest interviewer workload, which, together with other observations, led Groves and Magilavy to the interesting conclusion that the (Kish) survey model underlying ϱ^* might be further developed to reflect larger interviewer variability in the initial cases completed by the interviewers.

Like Collins and Kish in their studies, Groves and Magilavy did not find any evidence that factual items as a class are subject to different interviewer effects than are attitudinal questions. Groves and Magilavy also discuss two issues concerning interviewer effects which have been largely overlooked in the literature, namely, the stability of the estimates of ϱ^*, and the causes of interviewer effects.

4. Discussion

In this paper we have reviewed several important applications of early survey models. Some of them, like the U.S. Census Bureau model and the Kish model, have shown a broad applicability. They have been used not only within the agency for which they primarily were designed, but also in other contexts with quite different survey environments. Despite this, there is reason to ask why survey models have not been more extensively applied in survey quality work. After all, outside the United States, Canada, and perhaps some other countries, applications of survey models are rare. Applications do appear in certain experiments and surveys, but often as a result of a single researcher's interest in the field. These scattered applications often concern a small number of variables only. There are different reasons for this state of affairs.

i. The models do not cover all possible error sources. The survey models we reviewed in Section 2 mainly concern content errors and sampling errors. They do not account for, e.g., frame errors, coverage errors, and nonresponse errors.

ii. The models are based on assumptions that are seldom met in survey situations. For instance, when estimating the simple response variance, a common assumption is that reinterview responses are independent of the original answers and have the same distribution. Bailar and Dalenius (1969) showed that the simple response variance component could be estimated even if the reinterviews and the original interviews were permitted to be dependent. This, however, requires a second reinterview survey, which, for practical reasons, may be difficult and certainly expensive to implement. Another example is that the early models almost always presuppose a simple sampling design, whereas, in practice, survey designs are usually much more complex.

iii. The experimental designs necessary for estimating the components in the early models are expensive to implement. When personal interviews are used, interpenetration of the interviewers' workloads can be very costly if the study area is large. This problem can be diminished if the population under study and the population of interviewers are stratified and the model is applied in each stratum, as suggested by Sukhatme, or if the populations are grouped as assumed in the Hansen et al. (1951) model. However, even these designs would be expensive for organizations such

as Statistics Sweden where the interviewers are spread over the country and work alone in large areas. Another practical problem associated with interpenetration of interviewer assignments occurs in countries where the sampling units are individuals (and not housing units). Tracking respondents then becomes an important part of the interviewers' work. Since tracking respondents requires good knowledge of the local environment, interpenetrating could lead to increased nonresponse problems. In telephone interviews, the cost problems with interpenetration can almost be ignored, but the nonresponse problem cannot. Also, the costs of conducting reinterviews are considerable since large reinterview samples are needed for estimating the simple response variance component with an acceptable precision.

5. References

Abul-Ela, A. A., Greenberg, B. G., and Horvitz, D. G. (1967): A Multiproportions Randomized Response Model. Journal of the American Statistical Association, 62, pp. 990–1008.

Bailar, B. A. and Biemer, P. P. (1984): Some Methods for Evaluating Nonsampling Error in Household Censuses and Surveys. In Rao, P. S. R. S. and Sedransk, J. (eds.): W. G. Cochran's Impact on Statistics, pp. 253–274. John Wiley & Sons, New York.

Bailar, B. A. and Dalenius, T. (1969): Estimating the Response Variance Components of the U. S. Bureau of the Census' Survey Model. Sankhyā: Series B, pp. 341–360.

Casady, R. J. (1966): A Model for Analysis of Reinterview Data in the NCHS Health Interview Survey. Memo. U. S. Department of Health, Education, and Welfare.

Cochran, W. G. (1968): Errors of Measurement in Statistics. Technometrics, 10, pp. 637–666.

Collins, M. (1980): Interviewer Variability: A Review of the Problem. Methodological Working Paper No. 19. Social and Community Planning Research. London.

Deming, W. E. (1944): On Errors in Surveys. American Sociological Review, 9, pp. 359–369.

Des Raj (1968): Sampling Theory. McGraw-Hill, New York.

Fellegi, I. (1964): Response Variance and Its Estimation. Journal of the American Statistical Association, 59, pp. 1016–1041.

Forsman, G. (1987): Early Survey Models and Their Impact on Survey Quality Work. Proceedings of the U. S. Bureau of the Census' Third Annual Research Conference, Baltimore, MD, pp. 24–45.

Groves, R. M. and Magilavy, L. J. (1986): Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. Public Opinion Quarterly, 50, pp. 251–266.

Hansen, M. H., Hurwitz W. N., and Bershad, M. A. (1961): Measurement Errors in Censuses and Surveys. Bulletin of the International Statistical Institute, 38:2, pp. 359–374.

Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964): The Estimation and Interpenetration of Gross Differences and the Simple Response Variance. In Rao, C. R. (ed.): Contributions to Statistics. Statistical Publishing Society, Calcutta.

Hansen, M. H., Hurwitz, W. N., Marks, E. S., and Mauldin, W. P. (1951): Response Errors in Surveys. Journal of the American Statistical Association, 46, pp. 147–190.

Hansen, M. H. and Tepping, B. J. (1969): Progress and Problems in Survey Methods and Theory Illustrated by the Work of the United States Bureau of the Census. In Johnson, N. L., and Smith, H. (eds.): New Developments in Survey Sampling. Wiley-Interscience, New York, pp. 1–26.

Jabine, T. B. and Tepping, B. J. (1973): Controlling the Quality of Occupation and Industry Data. Bulletin of the International Statis-

tical Institute, 45, pp. 360–389.

Kish, L. (1962): Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, pp. 92–115.

Kish, L. and Lansing, J. B. (1954): Response Errors in Estimating the Value of Homes. Journal of the American Statistical Association, 49, pp. 520–538.

Krotki, K. P. and Hill, C. J. (1978): A Comparison of Correlated Response Variance Estimates Obtained in the 1961, 1971 and 1976 Censuses. Survey Methodology, 4, pp. 87–99.

Lyberg, I. (1986): Om evalveringsmetoder för Folk- och bostadsräkningen (FoB). Memo. Statistics Sweden. (In Swedish.)

Mahalanobis, P. C. (1946): Recent Experiments in Statistical Sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, 109, pp. 327–378.

Murthy, M. (1967): Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

Palmer, G. L. (1943): Factors in the Variability of Response in Enumerative Studies. Journal of the American Statistical Association, 38, pp. 143–152.

Rice, S. A. (1929): Contagious Bias in the Interview. American Journal of Sociology, 35, pp. 420–423.

Sanchez-Crespo, J. L. (1973): Balance Between Sampling and Non-sampling Errors in Spanish Official Statistics. Bulletin of the International Statistical Institute, 45, pp. 329–367.

Sanchez-Crespo, J. L. (1981): Spanish Experience on the Estimation of Some Components of the Total Error. Statistical Commission and Economic Commission for Europe. Conference of European Statisticians, June 1–4, 1981.

Stock, J. S. and Hochstim, J. R. (1951): A Method of Measuring Interviewer Variability. Public Opinion Quarterly, 15, pp. 322–334.

Strecker, H., Wiegert, R., and Kafka, K. (1984): Practical Determination of a Response Variance on the Basis on Survey Models with Reenumerations. Jahrbücher für Nationalökonomie und Statistik, 199, pp. 1–31.

Sukhatme, P. V. (1954): Sampling Theory of Surveys With Applications. Rome: United Nations Food and Agriculture Organization. Published by: The Iowa State College Press and The Indian Society of Agricultural Research.

Sukhatme, P. V. and Seth, G. R. (1952): Nonsampling Errors in Surveys. Journal of the Indian Society of Agricultural Statistics, 4, pp. 5–41.

Swensson, B. (1969): Alternativa surveymodeller för dikotoma variabler. Report no. 19 of the Errors in Surveys research project. Institute of Statistics. University of Stockholm. (In Swedish.)

Tepping, B. J. and Boland, K. L. (1972): Response Variance in the Current Population Survey. Working Paper No. 36, Bureau of the Census. U. S. Government Printing Office, Washington, D. C.

U. S. Bureau of the Census (1978): The Current Population Survey. Design and Methodology. Technical Paper 40. U. S. Government Printing Office, Washington, D. C.

U. S. Bureau of the Census (1985): Evaluating Censuses of Population and Housing. Statistical Training Document, ISP-TR-5, Washington, D. C.

Warner, S. L. (1965): A Randomized Response Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, pp. 63–69.

Zarkovich, S. S. (1966): Quality of Statistical Data. Rome: Food and Agriculture Organization of the United Nations.

Received May 1987
Revised January 1989

Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 57–67

Assessment of Regression Based Disclosure Risk in Statistical Databases

Michael A. Palley[1]

Abstract: Regression based disclosure can threaten the confidentiality of data stored in statistical databases. This holds true even when the database employs inference controls. This article describes database characteristics that correspond to high and low levels of risk of regression based disclosure. This leads to guidelines for the assessment of a statistical database's degree of risk. Identification of factors that contribute to disclosure risk is a first step in the development of new inference controls.

Key words: Confidentiality; statistical databases; disclosure; database management systems; security.

1. Introduction

The U.S. Census Bureau and other agencies collect data from individuals and later release the information in aggregate form. The respondent relies on the agency's promise of confidentiality when providing responses.

It is well known that promises of confidentiality are difficult to keep. Agencies that provide online statistical databases rely on various inference controls to maintain the confidentiality of collected data.

The literature discusses several inference controls for statistical databases. These include: limiting database responses to those with a minimum acceptable response set size; providing responses based on a random sample of the response set, (Denning (1980)); providing responses based on randomly perturbed (no bias) confidential data ("random data perturbation") (Beck (1980); Traub, Yemini, and Wozniakowski (1984)); and multidimensional transformation of the data matrix (Dalenius and Reiss (1982); Schlörer (1981)). An extensive discussion of the various existing inference controls is found in Chin and Ozsoyoglu (1982).

Inference controls offer only limited protection to the confidentiality of data in a statistical database. Note that restricting a statistical database's responses may conflict with the objective of a statistical database, i.e., to make aggregate statistical information available. Inference controls attempt to maintain the accuracy of database responses while protecting confidentiality.

When the confidentiality of data has been violated, "disclosure" has occurred. A method that can be employed to lead to the disclosure of confidential data is referred to as a "disclo-

[1] Associate Professor, Department of Statistics and Computer Information Systems, Baruch College – CUNY, 17 Lexington Avenue, Box 513, New York, N.Y. 10010, U.S.A.

sure technique." As will be discussed in this article, inference controls have only limited success in maintaining confidentiality when a regression based disclosure technique is applied to the database.

To understand the significance of our problem, it is necessary to further define statistical disclosure. Duncan and Lambert (1987) describe four ways to compromise the confidentiality of microdata that are considered in the literature.

Type I: Identification of a respondent from a released file (Duncan and Lambert call it an "identity disclosure"). (Spruill (1983); Paass (1985); Strudler, Oh, and Scheuren (1986).)

Type II: Obtaining information about a respondent by linking a record to the respondent ("attribute disclosure"). (Cox and Sande (1979).)

Type III: Gaining new information about a respondent from released data, even if no particular record is linked to the respondent, and even if the new information is inexact ("inferential disclosure"). (Dalenius (1974).)

Type IV: Disclosure of confidential information about a population or model, e.g., the relationship between employee characteristics and salary of a group (Palley and Simonoff (1986)). Duncan and Lambert call these "population disclosure" and "model disclosure" respectively.

Duncan and Lambert note that the tax compliance model of the U.S. Internal Revenue Service is a potential target for model disclosure. When a population has relatively few or insignificant outliers, individuals' confidential data can be estimated with these models, leading to inferential (Type III) disclosure as well. These issues are important for statistical agencies.

It has been shown (Palley and Simonoff (1987)) that a regression based technique can lead to inferential disclosure (Type III) as well as population or model disclosures (Type IV). The disclosure risk exists even when the online statistical database disallows the application of regression methodology. This disclosure technique has been shown effective even in the presence of existing inference controls. We identify characteristics that render a statistical database vulnerable to regression based disclosure.

We begin with a brief discussion of online statistical database systems. Section 3 summarizes the technique of regression based disclosure. Section 4 discusses factors that impede the use of the regression based disclosure technique. Section 5 presents assessment guidelines to determine a statistical database's risk of regression based disclosure. Finally, the impact on regression based disclosure control is presented in Section 6.

2. A Statistical Database

NAME	AGE	TITLE	YEARS	EDUC	DEPS	SALARY
JONES	32	VP	10	MBA	3	45 000
SMITH	40	AVP	10	BS	1	36 200
WATSON	27	MGR	2	MBA	0	37 500

Fig. 1. A section of a statistical database

A statistical database contains n records, each having m fields or "attributes" that contain values. Figure 1 presents a small section of a fictional statistical database. Some of these attributes (e.g., NAME) are unique record identifiers, called "keys" (Denning (1978)). Record values for keys are shielded by the database management system (DBMS) in order to prevent identification of described individuals. Other attributes (e.g., SALARY) contain confidential data. The attribute containing confidential data (referred to as the "confidential attribute") which the intruder seeks to disclose will be called "*Xc*."

The statistical database provides aggregate statistics to the user, for example MEAN SALARY = 35230, MEAN SALARY (WHERE AGE < 30) = 32000, etc. The parameters of the query, in our example AGE < 30, is referred to as the "characteristic" of the query (Denning (1978)). Note that a characteristic can involve several non-key attributes. The set of records that conform to a specific characteristic is referred to as the "response set." A statistical database is assumed to provide MEAN, COUNT (e.g., COUNT (where AGE = 30) = 35), and STANDARD DEVIATION (SD of SALARY (where AGE < 30) = 5010) query facilities. A simple technique to turn aggregate database responses into disclosure might be to identify records having unique characteristics. For example, the first record in Fig. 1 is the only record with AGE = 32 AND TITLE = VP AND YEARS = 10. Consequently, a typical inference control is to refuse to answer queries whose response set size is one, or relatively small.

3. Disclosure Using the Regression Based Technique

The technique of disclosing confidential data in a statistical database using regression methodology is developed, and described in detail in Palley (1986) and Palley and Simonoff (1987). Basically, a regression model is derived from the statistical database. The attribute storing confidential data serves as the dependent variable in the regression. Non-confidential, non-key attributes are candidates as predictor variables for the regression model. The model is built under the assumption that the DBMS precludes the direct application of regression methodology to the database. Hence the intruder must apply regression methodology using indirect means.

Based on the intruder's knowledge of predictor variable values (non-confidential "supplemental knowledge") relating to the individual whose confidential data is the "target," a statistical estimate of the target's confidential attribute value can be made. Disclosure of confidential data within a range of values constitutes inferential disclosure of the statistical database (notably Beck (1980); Traub et al. (1984); Denning (1978); Dalenius (1977); and Loynes (1979)). The following is a brief discussion of this technique.

3.1. The disclosure technique

Regression based disclosure involves three steps: selection of candidate predictor variables, generation of characteristic based queries, and derivation of what is called a "synthetic database." A synthetic database will exhibit regression relationships that mimic the statistical database. It is created using legitimate means, i.e., the technique does not violate any of the database's inference controls. Since the statistical database precludes application of regression methodology, the disclosure technique circumvents this control through derivation of a "synthetic" database.

3.1.1. Selection of candidate predictor variables

The technique begins with the selection of the confidential attribute of interest. Other useful

preliminary information is the list of database attributes and the number of records in the database. Henceforth, the statistical database will be referred to as the "actual database" (to be differentiated from the synthetic database). The intruder must then select a set of candidate predicator variables for the regression model from the non-key, non-confidential attributes. Selection is made on the basis of assumed attribute relationships or known correlations between attributes. For each candidate predictor variable, the intruder queries the actual database to create a histogram (i.e., a frequency distribution of values for each candidate predictor variable).

3.1.2. Generation of characteristic based queries

Once frequency tables are formed for each candidate predictor variable, the intruder randomly generates a value for each variable, based on the variable's frequency distribution. A combination of single values for each candidate predictor variable will constitute a characteristic. Each characteristic is used to generate three queries of the actual database, MEAN, COUNT, and STANDARD DEVIATION (SD). As an example, referring to Fig. 1, a possible characteristic might be (AGE = 35–40, TITLE = AVP, YEARS = 17, EDUC = BA, DEPS = 2). Let us call this characteristic P1. This characteristic is used to query the actual database: MEAN SALARY WHERE P1; SD SALARY WHERE P1; COUNT WHERE P1 (called F for frequency). The characteristic and responses to these queries are logged onto a table called the "interim tuple table (ITT)" (see Fig. 2). The strategy is repeated multiple times until an adequate percentage of database records are described. What constitutes an adequate percentage is a research issue and is discussed in Section 4.

/ AGE	/ TITLE	/ YEARS	/ EDUC	/ DEPS	/ SALARY	/ SD	/ F	/
/ 35–40	/ AVP	/ 17	/ BA	/ 2	/ 32 171	/ 5011	/ 3	/
/ 40–45	/ AVP	/ 15	/ MBA	/ 1	/ 42 131	/ 2019	/ 8	/
/								/
/								/

Fig. 2. Interim tuple table

3.1.3. Synthetic database derivation

The next stage of the technique is the derivation of the synthetic database. Once created, the synthetic database is available for regression analysis, or any other type of analysis that the intruder desires. Creation of the synthetic database from the ITT proceeds as follows. The pooled variance (s^2_{pooled}) of the ITT is derived, based on our standard deviation findings for each query response. Each ITT record is copied F times into the synthetic database. For each of the F copies, we vary the value for its confidential attribute by adding random normals distributed over (0, s_{pooled}) to its mean value (from the ITT). The pooled variance is utilized since we seek to simulate the overall variability of the actual database in the synthetic database.

Interim Tuple Table

Characteristic AGE/TITLE/YEARS/ ...	$\bar{X}c$	SD	F
35–40 AVP 17 ...	32 171	5 011	3

Synthetic Database

	Characteristic AGE/TITLE/YEARS ...	$\hat{X}c$
⇒	35–40 AVP 17 ...	29 315
⇒	35–40 AVP 17 ...	37 208
⇒	35–40 AVP 17 ...	35 211

Fig 3. Transformation into the synthetic database

The synthetic database is complete when all of the records in the ITT have been transformed in this way. Regression based disclosure now applies stepwise regression analysis directly to the synthetic database. This regression model, referred to as a "disclosure model," is used to estimate values for the confidential attribute. The estimate is based on the intruder's supplemental knowledge of the non-confidential attribute values for the target individual.

It should be noted that despite the seeming complexity of the approach, the technique could be applied rather easily with the use of a microcomputer. Characteristic generation, and logging the responses onto the ITT occur independently of the actual statistical database. The only points of contact between the intruder and the actual database are the initial building of frequency distributions, and the characteristic based querying of the database. All other functions could be performed on a stand-alone microcomputer at the intruder's convenience.

3.2. The technique as a threat to confidentiality

Palley (1986) and Palley and Simonoff (1987) reported the results of validation of this technique. Regression based disclosure was attempted on several subsamples of the 1980 U.S. Census Microdatabase C-sample for the State of New York. Each of these subsamples was treated as a statistical database. The attribute FAMILY INCOME was considered to be confidential. Characteristic-based queries were applied to the actual databases using our technique. The specific findings are lengthy, and are recounted in detail in Palley and Simonoff (1987).

The analysis considered the following criteria. (a) The degree to which the synthetic database resembled the actual database. This was determined by cross-validating the synthetic database derived disclosure model against the actual database. (b) The quality of estimates of confidential data produced by our disclosure model. Regression based disclosure was found in our analysis to be an effective threat to statistical database confidentiality.

Palley and Simonoff (1987) found that the regression based disclosure technique performed well even in the presence of inference controls. It was shown theoretically that in situations where the "random sample queries" (Denning (1980)) and multidimensional transformation (Dalenius and Reiss (1982)) controls were employed, there was no effect on the performance of the regression based disclosure technique.

It was demonstrated empirically that the control of refusing to answer queries with small response sets had no significant effect on regression based disclosure. This held true until the minimum response set size was set to a threshold level, relative to the size of the database. It was also shown that at the same threshold level, accurate aggregate statistics would be withheld from legitimate users. Therefore, in reality the control offered no protection. Finally, random data perturbation was empirically shown to have little effect on the performance of the regression based disclosure technique. In fact, the research derived an adjustment factor to filter any bias that the perturbation may have added to the synthetic database (assuming the perturbation level of the actual database is known).

These controls failed since they preserve the overall statistical characteristics of the database. These controls must permit the database to answer queries without significantly distorting the responses. If responses are significantly distorted, the database will fail to provide accurate data to legitimate users. As long as accurate responses are provided by the database, regression based disclosure can occur, regardless of these inference controls.

4. Complicators of the Regression Based Approach

This research identifies factors that complicate disclosure of confidential data in a statistical database using the regression based technique. Knowledge of these factors will assist us in the evaluation of the disclosure risk of statistical databases. Future research based on this work may suggest new inference controls that can deter regression based disclosure.

Disclosure using the regression based technique requires the existence of a regression relationship in the actual database. This regression relationship must significantly describe the confidential attribute. The lack of such a statistical relationship will render the disclosure technique harmless. We proceed while assuming the existence of this relationship.

The major factors that complicate disclosure using the regression based technique are: (a) combinatorial explosion of possible characteristics, (b) uniform distribution of actual database records corresponding to the possible characteristics, and (c) minimum response set size that is large relative to the actual database. All of these factors eventually lead to an ITT that describes little of the actual database. Creation of a synthetic database from such an ITT will result in a synthetic database, and consequently a disclosure model, that bears little resemblance to the actual database. These factors are now discussed in detail.

4.1. Combinatorial explosion of possible characteristics

Regression based disclosure begins with the querying of the actual database to build histograms of candidate predictor variables. The next step is the random generation of characteristics used for querying the actual database. Combinatorial explosion, used here, is the presence of a large number of possible characteristics relative to the actual database size. For example, if a characteristic consists of attributes AGE (perhaps 50 possible values), TITLE (10 possible values), and YEARS-WITH-FIRM (30 possible values), then there would be 15 000 potential characteristics. The number of combinations worsens drastically with each additional attribute being used in a characteristic.

When there are a large number of possible characteristics relative to the actual database size, few database records (i.e., individuals described in the database) will conform to any given characteristic, hence small response set sizes. If an inference control that prevents responses to queries with small response sets is employed, many of these queries to the actual database will go unanswered. Even if the small response set queries are answered, the marginal value of asking these queries is relatively low. Considering the risk of detection related to asking many queries, the combinatorial explosion problem is potentially detrimental to the regression based disclosure technique.

There are two possible causes of combinatorial explosion. The first is the presence of many predictor variables in the regression model that exists in the actual database. An intruder's strategy would be to utilize best subset regression in order to limit the number of variables in the model. However, when a regression model has many predictor variables that each contribute relatively little to estimating the confidential attribute (measured by R^2), best subset regression would not be helpful. In addition to combinatorial explosion, a large set of predic-

tor variables requires that the intruder have a great deal of non-confidential data (supplemental knowledge) about the target in order to exploit the disclosure model. The more supplemental knowledge missing, the less useful the disclosure model will be for inferring an individual's confidential data.

Wide domains of predictor variable values would also contribute to combinatorial explosion. An intruder's strategy to remedy this would be to cluster values into subsets, e.g., AGE: 35–40 ... However, some continuous variables may have wide value ranges that will not form meaningful clusters.

4.2. *Interim tuple table insufficiency*

Combinatorial explosion will lead to an insufficient ITT. By insufficient, we mean that the ITT accounts for a small number of actual database records relative to the actual database size.

DATABASE SIZE	374					752				
PCTG OF RECORDS ACCOUNTED IN ITT	63%	50%	30%	20%		68%	61%	52%	47%	41%
PREDICTIVE R^2	.47	.42	.42	−.05		.44	.42	.42	.39	.25

Fig. 4. Relationship of ITT records to predictive R^2

Figure 4 emanates from the Palley and Simonoff (1987) study of two U.S. census subsamples. It is presented for the first time here. As the regression based disclosure technique was applied, the problem of ITT insufficiency was demonstrated. Various ITTs were created from each of two statistical databases. The number of possible characteristics (number of attributes in a characteristic; and number of possible values for each of the attributes in a characteristic) varied in the creation of each of these ITTs. The different number of possible characteristics led to ITTs that accounted for different percentages of records from their respective actual database (middle row of Fig. 4). These ITTs were employed to create disclosure models. The last row of Fig. 4 indicates the quality of the disclosure model derived from each synthetic database, as applied to the actual database, measured as predictive R^2.

Predictive R^2 is a measure of the fit of a regression model created on one set of data as applied to another. The formulation for predictive R^2 is:

$$1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares}}$$

A "perfect" disclosure model would have no residual sum of squares, and therefore have a predictive R^2 of 1. Since this disclosure model is being applied to a different set of data than it was created on, predictive R^2 can potentially be negative (as seen in Fig. 4). This would be true if the disclosure model is a worse estimator of confidential values than the sample mean.

Figure 4 indicates a relationship between the percentage of database records described in the ITT and the quality of a derived disclosure model. We observe a decline in quality of the disclosure model (measured by the predictive

R^2), as the percentage of actual database records described in the ITT declines.

Part of the reason that ITT insufficiency leads to a poor disclosure model is that an insufficient ITT leads to a synthetic database that has low variability of the confidential attribute. Those ITTs that described relatively few actual database records led to disclosure models that described their synthetic database extremely well ($R^2 \geq 0.8$). However these disclosure models were very poor descriptors of the actual database, hence they had no utility to the intruder.

Two other factors may result in an insufficient interim tuple table, namely uniform distribution of actual database records corresponding to the possible characteristics, and a large minimum response set size control.

4.3. Uniformly distributed characteristic distributions

Disclosure through this technique presupposes that individuals described in the database tend to cluster among a relatively limited subset of the possible characteristics. The regression based technique acts to capture those characteristics that describe a large proportion of the records in the database. For example, let us assume that there are ten thousand records in a statistical database. We will also assume that there are one thousand potential characteristics. Regression based disclosure works relatively well if the database records cluster non-uniformly among a subset of those characteristics. However, if the records cluster uniformly among most of the characteristics, it will take a prohibitive number of database queries in order to build a sufficient ITT. Furthermore, if there is a uniform distribution, the typical response set size will be relatively small. This could cause problems if there is a minimum response set size.

The intruder's strategy would be analogous to the strategy for combinatorial explosion: to reduce the number of possible characteristics (i.e., by reducing the number of attributes comprising a characteristic, or by grouping characteristic attribute values), hoping that distributions among these fewer combinations of characteristics are less uniform. However, if database records remain relatively uniformly distributed among the new possible characteristics, the intruder will not be able to solve this problem.

4.4. High minimum response set size

If a statistical database employs an inference control that refuses queries with a minimum response set size that is large relative to the database size, many queries will go unanswered. This will result in ITT insufficiency. It is noted that the strategy of raising minimum response set size past a point is a "two-edged sword." The strategy will protect against regression based disclosure, but only at the expense of failing to provide the legitimate user with useful aggregate statistics.

5. Risk Assessment Guidelines

At this stage, we seek to assess the risk of regression based disclosure for a statistical database. The assessment guidelines generally parallel our discussion of the complicators of disclosure. A diagram of factors that contribute to the risk of regression based disclosure is presented in Fig. 5.

Here risk is described qualitatively. There currently exists no means of quantifying the relative risk level. It is proposed that agencies that provide online statistical database facilities can assess the level of regression based disclosure risk by answering the following.

A. Does a regression relationship exist in the database, with a confidential attribute acting as dependent variable?

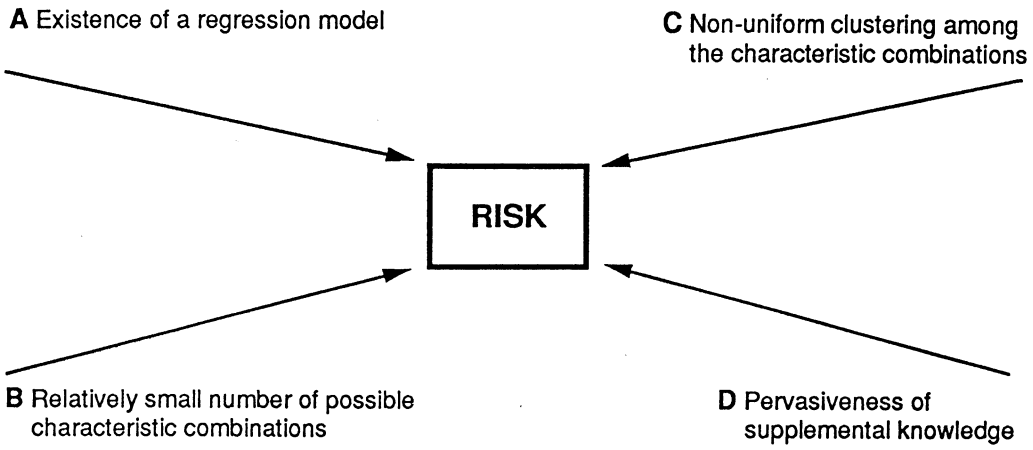


Fig. 5. Assessment model

Upon identification of those attributes which contain confidential data, the agency should perform correlational analysis to identify candidate non-confidential, predictor variables. Next, perform stepwise regression on the database. Existence of a sufficiently high R^2 regression model (sufficiency to be determined by the agency), and pervasiveness of supplemental knowledge (data for non-confidential attributes for individuals in the database) would be indicators of disclosure risk.

B. Is there a small number of characteristics relative to the database size?

B.1. Does the regression model require few predictor variables?

A minimum "safe" number of predictor variables is a function of the size of the database. As a general rule, the larger the database, the more predictor variables would be necessary to cause ITT insufficiency. In addition, a large number of predictor variables places an added burden on the intruder for extensive supplemental knowledge. The fewer predictor variables necessary for a disclosure model, the more at risk the statistical database. Note, as discussed, combinatorial explosion, based on too many predictor variables in a disclosure model, can be remedied by various intruder strategies. To assess risk under these strategies, an agency can assess the quality of regression models (in terms of R^2) that involve fewer predictor variables.This can be facilitated with best subset regression.

B.2. Do candidate predictor variables have few possible values?

The larger the domain of predictor variable values, the more difficult it is to create a disclosure model. A large number of predictor variable values will lead to a large set of possible characteristics, contributing to ITT insufficiency. Again, this is defined relative to database size. The ability to cluster wide variable value spreads into ranges is an effective way for the intruder to counter the combinatorial explosion problem. In order to assess the risk of this, an agency might try recording values into ranges, and test if the disclosure model remains relatively effective.

C. Are distributions of records among characteristics non-uniform?

The less uniform the distribution of database records among characteristics, the more regression based disclosure is facilitated. Uniform distributions result in query responses with small counts, making it difficult to adequately describe a large portion of the actual database with a reasonable number of queries (ITT insufficiency). An insufficient ITT leads to little variability of the confidential attribute data in the synthetic database, and likewise to a disclosure model that fails to describe the actual database. A database whose records cluster among relatively few characteristics has greater risk of regression based disclosure.

D. Is non-confidential data for the regression model's predictor variables generally available?

Inferential disclosure of a statistical database requires the intruder's knowledge of his target's values for predictor variables (supplemental knowledge). Alternative strategies are available to an intruder who has incomplete supplemental knowledge. The intruder may create a disclosure model involving only those predictor variables for which he has supplemental knowledge. Another strategy would be for the intruder to estimate missing values. Palley and Simonoff (1987) found that when supplemental knowledge was lacking for one or two (out of five) predictor variable values, an intruder could still perform inferential disclosure effectively. However, the higher a given predictor variable's t-value (measure of that variable's contribution to the regression model) in the disclosure model, the more impact its value's absence. Clearly, the more available the non-confidential data, the more disclosure risk.

These disclosure risk criteria are not necessarily exhaustive. It is possible that future research will yet determine other risk criteria.

6. Final Remarks

The notions of population disclosure and model disclosure run counter to the perceived goals of statistical agencies. Statistical agencies make information, and therefore, regression relationships publicly available. Legitimate users have a need for information. However, regression based disclosure can turn seemingly benign types of information into breaches of confidentiality. This is particularly a problem when a model disclosure is parleyed into inferential disclosure of an individual's information. This is a problem posed by the regression based disclosure technique.

A regression based technique has been found to defy existing inference controls. We have identified some critical factors that influence the risk posed by regression based disclosure. The reduction of statistical database disclosure risk has been investigated by, among others, Duncan and Lambert (1986 and 1987); Cox and Sande (1979); Dalenius and Reiss (1982); Traub et al. (1984). Nevertheless, the existing research does not specifically address disclosure risk posed by regression based techniques. Drastic controls, such as refusing to provide standard deviation responses, would also reduce the utility of the statistical database to legitimate users. Replacing standard deviation responses with minimum and maximum bounds, besides limiting the information available to legitimate users, would also pose new disclosure risks. Consequently, there are no simple solutions.

Nevertheless, identification of the critical factors in regression based disclosure is a step in the development of further controls. Future research will continue to address these problems. In the interim, this research highlights some limitations in our ability to protect the confidentiality of collected statistical data in online databases.

7. References

Beck, L. L. (1980): A Security Mechanism for Statistical Databases. ACM Transactions on Database Systems, 5, pp. 316–338.

Chin, F. Y. and Ozsoyoglu, G. (1982): Auditing and Inference Control in Statistical Databases. IEEE Transactions on Software Engineering, SE-8, 6, pp. 574–582.

Cox, L. H. and Sande, G. (1979): Techniques for Preserving Statistical Confidentiality. Proceedings of the 42nd Meeting of the International Statistical Institute, Manila, December, 1979.

Dalenius, T. (1974): The Invasion of Privacy Problem and Statistics Production – An Overview. Statistisk tidskrift, 12, pp. 213–225.

Dalenius, T. (1977): Towards a Methodology for Statistical Disclosure Control. Statistisk tidskrift, 15, pp. 429–444.

Dalenius, T. and Reiss, S. (1982): Data-Swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference, 6, pp. 73–85.

Denning, D. (1978): A Review of the Research of Statistical Database Security. In Foundations of Secure Computation, edited by R. DeMillo et al., Academic Press, New York.

Denning, D. (1980): Secure Statistical Databases with Random Sample Queries. ACM Transactions on Database Systems, 5, pp. 291–315.

Duncan, G. and Lambert, D. (1986): Disclosure-Limited Data Dissemination. Journal of the American Statistical Association, 81, pp. 10–18.

Duncan, G. and Lambert, D. (1987): The Risk of Disclosure for Microdata. Statistics of Income and Related Administrative Record Research: 1986–1987. Internal Revenue Service Publication No. 1299, Government Printing Office, Washington, D.C., pp. 325–332.

Loynes, R. M. (1979): Discussion of the Papers by Professor Dalenius and Professor Durban. Journal of the Royal Statistical Society, Series A, 142, pp. 325–326.

Paass, G. (1985): Disclosure Risk and Disclosure Avoidance for Microdata. Paper presented at International Association for Social Service Information and Technology, May 1985.

Palley, M. A. (1986): Security of Statistical Databases: Compromise Through Attribute Correlational Modeling. Proceedings of the Second International Conference on Data Engineering, IEEE Computer Society, Washington, D.C., pp. 67–74.

Palley, M. A. and Simonoff, J. S. (1986): Regression Methodology Based Disclosure of a Statistical Database. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 382–387.

Palley, M. A. and Simonoff, J. S. (1987): The Use of Regression Methodology for the Compromise of Confidential Information in a Statistical Database. ACM Transactions on Database Systems, 12, pp. 593–608.

Schlörer, J. (1981): Security of Statistical Databases: Multidimensional Transformation. ACM Transactions on Database Systems, 6, pp. 95–112.

Spruill, N. (1983): The Confidentiality and Usefulness of Masked Business Microdata. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 602–607.

Strudler, M., Oh. H. L., and Scheuren, F. (1986): Protection of Taxpayer Confidentiality with Respect to the Tax Model. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 375–381.

Traub, J. F., Yemini, Y., and Wozniakowski, H. (1984): The Statistical Security of a Statistical Database. ACM Transactions on Database Systems, 9, pp. 672–679.

Received September 1987
Revised March 1989

Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 69–81

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

Recent Work With Microcomputers for Census Processing in Developing Countries

Vivian Toro and Kathleen Chamberlain[1]

Abstract: This paper is a revision of the paper entitled "The Use of Microcomputers for Census Processing in Developing Countries" written by Vivian Toro and Thomas Melaney and presented at the American Statistical Association meetings in August 1987. The authors discuss how microcomputers offer potential solutions to many of the problems developing countries encounter when processing census data with mainframe computers. The authors also describe the Integrated Microcomputer Processing System (IMPS) developed by the International Statistical Programs Center (ISPC), U. S. Bureau of the Census, and present three case studies of the use of IMPS by statistical offices in developing countries.

The experiences of a number of developing countries confirm that microcomputers are technologically sound and cost-effective tools for processing censuses. The continuing technological advancement and refinement of microcomputer software make their use even more advantageous.

This paper focuses on the recent experience of Burkina Faso and Senegal in using microcomputers. A summary is given of the use of IMPS by an increasing number of countries. Additionally, this paper describes the latest enhancements and future plans for IMPS. It concludes with some thoughts on the use of microcomputers in the areas of data collection and processing, as well as in the use and dissemination of census data.

Key words: Microcomputers; census; developing countries; software; data processing; data collection; integrated.

[1] International Statistical Programs Center, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are those of the authors and do not necessarily reflect those of the Census Bureau. An earlier version of this paper was presented at Symposium 88: The Impact of High Technology on Survey Taking. Ottawa, Canada, October 1988.

1. Microcomputers as Tools for Processing Census Data

Technological developments over the last two years have made the processing of large volumes of data on microcomputers even more practical. Problems normally associated with large-scale processing jobs such as population censuses are quickly disappearing. Data stor-

age devices for microcomputers are more plentiful, faster, more reliable, smaller in size, higher in capacity, and less expensive. Processing speeds of microcomputers continue to increase and memory continues to expand.

Statistical offices in developing countries have found this technology easily accessible. For the most part, finding a local microcomputer vendor is no longer a problem. Typically, three or more different vendors are likely to be represented. This more competitive market has not only expanded the number of options, but also has reduced the prices and improved the terms of maintenance contracts.

Microcomputers are now an integral part of the operation of most statistical offices in developing countries. Microcomputers are being used for processing surveys, scheduling, budgeting, word processing, and other activities. A cadre of microcomputer literates exists in most of these countries. In addition, a variety of microcomputer software packages in different languages is now available.

Instead of asking whether or not to use microcomputers, statistical offices are now wondering how best to use them. They are asking what type of microcomputer software and hardware will best meet their short- and long-term needs. They are interested in obtaining the resources to buy this hardware and software, to train their staff, to provide hardware maintenance, and to integrate microcomputers into existing mainframe environments.[2]

As statistical offices have become more comfortable with the microcomputer technology, many have chosen to use microcomputers for processing all or part of their census data. Burkina Faso, Senegal, and Micronesia were some of the pioneers that opted to process their census data using microcomputers. They were followed by countries such as Niger, Swaziland, Yemen, Honduras, Uruguay, and the Central African Republic. Countries including the Philippines, Pakistan, and Côte d'Ivoire chose to use microcomputers just for the entry of the data from census questionnaires. Countries such as Tanzania and Cameroon chose to share processing between microcomputers and a mainframe computer.

Microcomputers have provided a solution to some of the problems encountered in the past by countries which used only mainframe computers to process their census data. Such countries have been able to afford the computing capability of the microcomputer within their limited budgets. The availability of several microcomputers has resolved most of the computer accessibility problems. Microcomputer maintenance is easier, and environmental standards are less restrictive. Additionally, training requirements are less demanding given the availability of software that is easier to learn and use, even for persons with little background in data processing.

The availability of microcomputer software which addresses the particular needs of census data processing has contributed greatly to the success microcomputers have had in statistical offices. Most of the countries mentioned above have used or plan to use IMPS.

2. The Integrated Microcomputer Processing System

2.1. *Background*

The introduction of more powerful microcomputers in the early 1980s was not accompanied by microcomputer software which adequately addressed the requirements of census processing. ISPC embarked on the development of IMPS to provide microcomputer software for the major tasks of census data processing: data entry, editing, tabulation, analysis, and operational control. Because a project of this magnitude takes considerable time and resources, ISPC decided to use existing software where

[2]In this paper, any reference to mainframe computers will also include minicomputers.

possible, and to provide intermediate products throughout the development of IMPS.

The design objectives of IMPS included ease of use, availability of a common data dictionary, ability to run with a standard microcomputer configuration, and modularity. The concept of modularity was of particular importance in the early development of IMPS. The system was designed in such a way that modules could be easily swapped as software improvements were made. For example, prior to the development of the IMPS data entry component, CENTRY (Census data ENTRY), commercial data entry packages were commonly used. The substitution of CENTRY does not require modification of the other application modules such as editing and tabulation. Another advantage of modularity is flexibility. The user can choose which tasks to perform using IMPS and which to perform using other software or other hardware.

The first step in the development of IMPS was to make the software packages CONCOR and CENTS 4 available on microcomputers. CONCOR (CONsistency and CORrection) and CENTS 4 (CENsus Tabulation System version 4) are editing and tabulation packages, respectively, developed by ISPC. These packages run on most mainframe computers and were in use by over 100 developing-country statistical offices by 1985. Both packages were written in COBOL, to ease the job of adapting them to the number of different mainframe computers found in developing countries. The availability of Realia COBOL, a powerful and efficient COBOL compiler for microcomputers, facilitated the adaptation of CONCOR and CENTS 4 to the microcomputer environment.

The first microcomputer versions of CONCOR and CENTS 4 appeared in 1984. The user instruction sets for these versions were the same as for the mainframe versions. This meant that the same CONCOR or CENTS application programs could run on either the mainframe computer or on microcomputers at statistical offices where both types of hardware were available. A number of statistical offices, including Senegal, Cameroon, and Somalia, took advantage of this flexibility.

Data capture, often a major bottleneck in census processing, was the next hurdle. ISPC evaluated commercial data entry packages to identify the best software for the capture of census and survey data. The evaluation identified two packages, ENTRYPOINT from Datalex, Inc., and RODE/PC from DPX, Inc., as meeting the performance criteria. At the time of the evaluation, RODE/PC had a slight advantage over ENTRYPOINT in terms of cost and speed. Therefore, ISPC recommended RODE/PC for census data entry on microcomputers.

The widely used Computer Programs for Demographic Analysis (CPDA), written by the Bureau of the Census in the early 1970s, was made available for microcomputers by Westinghouse Public Applied Systems under the sponsorship of the U.S. Agency for International Development (USAID). The Microcomputer Programs for Demographic Analysis (MCPDA) is one of several demographic analysis packages available on microcomputers. Currently, the Bureau of the Census is developing a more comprehensive package called Model Spreadsheets for Demographic Analysis (MSDA). Additionally, it is preparing a manual on the analysis of census data. This manual brings together the most useful methods of demographic analysis and projection, including brief descriptions, discussions of the advantages and disadvantages of different methods, and other helpful hints for interpreting the results. The manual will also discuss available software for performing each type of analysis. The MSDA will be available in about one year and the manual during the following year.

The statistical analysis package PC-CARP, developed by Iowa State University, is the statistical analysis component of IMPS. Over the

past two years it has become widely used by statistical offices in developing countries for the calculation of variances in surveys, including post enumeration surveys. Users of PC-CARP include Zambia, Peru, Egypt, Zimbabwe, Costa Rica, Haiti, and the Philippines. The user interface of PC-CARP facilitates its use by analysts with little or no data processing experience.

With microcomputer software available for data entry, editing, tabulation, demographic analysis and statistical analysis, and with access to high capacity storage devices such as Bernoulli Boxes,[3] it became feasible to recommend census processing on microcomputers for countries with populations of 10 million or less. The 1985 census of Burkina Faso was the first major census to be processed entirely on microcomputers. RODE/PC, CONCOR, CENTS 4, and MCPDA were used. Burkina Faso's experience is described in detail in Section 3.

Since 1985, ISPC has added new features to IMPS. These include a common data dictionary, a more economical data entry module, improvements to CONCOR and CENTS, a census management and reporting system, and a census resource planning software.

[3]Bernoulli Boxes are removable cartridge-based hard-disk systems. They can serve as both mass-storage devices and backup systems. Information is stored in greater densities and higher data access speeds are allowed. Each cartridge unit has one or two cartridge readers. Depending on the model, the units can read data from either 10 or 20 megabytes cartridges. Since these cartridges are removable, they provide virtually unlimited capacity. The newer models also have up to 80 megabytes of hard-disk storage in addition to the removable cartridges. In addition, they are physically very resistant to rough treatment which would otherwise have destroyed or caused the loss of data on hard disks. They have been used for several years in developing countries and have proved to be highly reliable.

2.2. *New features*

2.2.1. Common data dictionary

CONCOR and CENTS 4 originally were developed as stand-alone packages for mainframe computers. During their development, little attention was paid to integrating the two packages because the major concerns were portability and use of the software on computers with very limited memory capacities. As a result, the user was forced to define the data file for the CONCOR application program, then again for the CENTS 4 application program, using a different notation each time. After CONCOR and CENTS 4 were adapted to the microcomputer, the data definition became even more cumbersome because commercial data entry packages such as RODE/PC required yet a third definition of the data. Multiple data definitions were not only time-consuming to code but were also error-prone.

The task of integrating CONCOR and CENTS through a common data dictionary resulted in CONCOR version 3 and CENTS version 5 which were released in early 1988. With these versions, the definition of the data file to be edited and tabulated is done only once. The interactive module of IMPS called DATADICT prompts the user for a name and the characteristics of each data item. The resulting definition, the Data Dictionary, is used by both CONCOR and CENTS.

In these new versions of CONCOR and CENTS, the user interface has been greatly simplified, and the CONCOR edit reports were made more concise. Several new features were added to CONCOR such as the ability to reference repeating data items and greater flexibility in creation of the extract files.

2.2.2 CENTRY: Software for data entry

As the cost of commercial data entry packages rose, it became apparent that statistical offices in developing countries needed a less expensive

alternative. For example, in order to enter questionnaires within one year from a country with a population of 10 million would require approximately 30 microcomputers, meaning 30 copies of the data entry software package. At about $600 (U.S.) per copy, data entry software becomes quite expensive. Moreover, RODE/PC and ENTRYPOINT contain features, such as extensive logic checks, that are essential for survey data entry but not needed for census data entry. For census data entry, speed is of primary importance since the volume of data is so great. Capturing what is on the questionnaire, consistent or inconsistent, is essential. Inconsistencies can be corrected later through CONCOR and should not be left up to data entry staff to correct. Most developing countries have trouble just attaining acceptable keying rates with systems that have no or very limited logic checks.

Keeping in mind the particular needs for the entry of census data, ISPC developed a package called CENTRY as the data entry module of IMPS. CENTRY is a screen-oriented menu-driven package which allows for developing data entry applications, entering and verifying data, and collecting statistics on data entry operations. CENTRY uses the same Data Dictionary as CONCOR and CENTS. Features of CENTRY which make it attractive for census data entry include: programmable data entry screens, valid value checking, automatic duplication of fields, cursor control, skip pattern control, record retrieval and modification, and operator statistics. The data entered through CENTRY are stored in an ASCII file ready for use by other packages such as CONCOR. The user-friendliness of the IMPS data dictionary and CENTRY allows a person with minimal training who is familiar with the census questionnaire to set up a data entry application in a day. CENTRY requires 256 kilobytes of memory and two floppy disk drives. It is written in the C language to promote fast execution.

CENTRY will be released in November 1988. Already several countries, including Central African Republic, Burundi, and the Philippines, are planning to use it for census data entry.

2.2.3. dCONTROL: Census management system

With a large data processing operation such as a national census, it is important to track each unit of data through the various processing phases. dCONTROL is an interactive census management and control system that helps census managers to monitor these phases. It also serves as a "check-in" facility for data by geographical unit, such as enumeration area (EA), thereby preventing duplication or omission of EAs. dCONTROL also allows for the production of preliminary count reports and other management reports.

In its current state, dCONTROL is a prototype that can be adapted to meet the particular needs of a country. It is being used in both Niger and Senegal where their statistical office data processing staff customized it to suit their specific needs. ISPC currently is developing a generalized version of dCONTROL which will be available in early 1989 in English, French, and Spanish.

dCONTROL requires a good cartographic system that accurately defines all the geographic and administrative boundaries for the census. The cartographic system serves as the basis for a geographic coding scheme that identifies each statistical area down to the EA level. dCONTROL monitors the various phases of processing at the EA level. The first phase is usually the receipt of the questionnaires from the field by the central office; the last phase is the keying of the census data.

dCONTROL is written in dBASE III PLUS programming language compiled using CLIPPER by Nantucket. It runs on an IBM PC/XT or PC/AT or compatible with 512 kilobytes of memory and at least 10 megabytes of auxiliary

storage, usually a hard disk. The amount of hard disk needed depends upon the number of EAs in the country. For a country of 8000 EAs, two megabytes of hard disk are sufficient to store the dCONTROL programs and the data.

2.2.4. CENPLAN: (CENsus PLANning)

Although great strides have been made in a very short time in the development of census processing software for microcomputers, there is still room for improvement. The goal is to make census results available as soon as possible after the census is taken. A number of factors contribute to the delay of census results, not the least of which is poor planning.

ISPC is developing a software package called CENPLAN which allows the census planner to determine the resources needed for census processing, given certain parameters such as population size, available computer equipment, and time constraints. Although CENPLAN cannot ensure that proper planning is done in time for a census, it provides the facilities to plan. CENPLAN uses spreadsheet software and will be available in French and Spanish, as well as English. The first release, which will address only the computer processing of a census, will be available in November 1988.

3. Case Studies

The number of countries using IMPS for the processing of census data continues to grow. Table 1 is a partial list of countries that have used, are using, or plan to use IMPS to process their census data. Statistical offices in developing countries are taking advantage of the system modularity that IMPS offers.

Some countries, including Niger, Senegal, Benin, Burkina Faso, the Central African Republic, Micronesia, Burundi, Yemen Arab Republic, Mali, Malawi, Swaziland, and the Pacific outlying areas of the U.S. have used or plan to use microcomputers for all aspects of their census data processing. Senegal, Yemen Arab Republic, and Mali, had originally planned to use microcomputers for data entry and a mainframe computer for further processing of the data. However, when they realized they could obtain results in a more timely and cost-effective manner, they decided to process the data on microcomputers. For example, they found that CONCOR and CENTS generally execute faster on the microcomputers than on their mainframes. Even when the actual execution time was theoretically the same or slightly faster on the mainframe computer, they found they could obtain results faster on the dedicated microcomputers because they did not have to share computer resources with other users.

Other countries such as Somalia and Côte d'Ivoire have developed their CONCOR and CENTS programs using microcomputers but are running these programs against data on mainframe computers. Some countries like Ghana and Ethiopia did their data entry on microcomputers, then uploaded the data to a mainframe computer for processing. Furthermore, countries like Honduras, the Philippines, and Tanzania are still trying to decide whether to use microcomputers, a mainframe computer, or a combination of the two for processing the census data. The selection criteria are not always solely technical. Some countries are committed to using mainframe computers for processing census data because their governments or donor agencies had purchased or procured the mainframe equipment several years ago with the express objective of using it for the census. Despite a technological revolution, amortization of this equipment was necessary before decision-makers could justify the purchase of new equipment. IMPS modular design allows statistical offices in developing countries to take advantage of existing hardware and software that is appropriate for census data entry. For example, Tanzania, and Cameroon

are using minicomputers such as ICL/DRS 300 and IBM System/36 for data entry. On the other hand, Yemen Arab Republic is using microcomputers for data entry but has contracted out the development of their data entry programs because specialized Arabic data entry is needed. All these countries plan to use or are using CONCOR and CENTS for the editing and tabulation of the data.

Additionally, as other IMPS components become available, statistical offices in developing countries can plan to take full advantage of it. For example, Burundi, the Central African Republic, and the Philippines are planning to use CENTRY for the entry of their census data.

Although designed primarily for census processing, IMPS also is being used for survey processing. Gambia, Morocco, Zambia, Senegal, Portugal, Ghana, Rwanda, Egypt, and American Samoa are using one or more of IMPS modules for surveys.

The following two case studies, Burkina Faso and Senegal, were presented in the Toro-Melaney paper as representative of the benefits of IMPS for census processing activities. A follow-up of their experience follows. The Federated States of Micronesia (FSM), the third case study presented in the Toro-Melaney paper, have continued to use microcomputers successfully to process the census data for some of the states. FSM is currently discussing the establishment of a national census that would standardize the date, questionnaire, and processing activities used by all the states.

3.1. Burkina Faso

In December 1985, the National Institute of Statistics and Demography (INSD) of Burkina Faso, formerly Upper Volta, conducted their second national census. It took them only 18 months to enter and process the data of their estimated population of 9 million.

Three IBM PC/AT microcomputers with 512 kilobytes of memory, and 20 megabytes of hard disk storage were used to process the data. The data entry equipment consisted of twenty-two IBM PC microcomputers each having 512 kilobytes of memory, and two IBM PC/XT microcomputers with 10 megabytes of hard disk storage and 640 kilobytes of memory. Related peripherals included six printers, two DIGIDATA tape drive units, and three Bernoulli Boxes with 10 megabyte removable cartridges. The tape drive units were used for backup and archival purposes, while the Bernoulli Boxes were used for primary storage of the census data. RODE/PC, CONCOR, and CENTS 4 were the software packages used for data entry, editing, and tabulation, respectively.

The INSD data processing staff for the census project consisted of two full-time programmers and a long-term United Nations advisor. In addition, the U.S. Bureau of the Census under an agreement with USAID, provided data processing assistance in the form of training, program development, and monitoring of the RODE/PC data entry application program, the CONCOR editing program, and the CENTS 4 tabulation programs.

The preliminary counts based on the manual counts were available by March 1986. Two shifts of 25 operators began keying the 15 % sample of the data in September 1986. Two months later, this data capture operation was complete, and by March 1987, the tabulations for the 15 % sample had been produced. The edited sample data occupy 11 Bernoulli cartridges of 10 megabytes each. (It should be noted that data entry was begun several months late due to delayed arrival of microcomputer equipment.)

The entry of the remainder of the data took place between December 1986 and May 1987. The 40 tables for the entire country were available by March 1988, 18 months after data entry began. The keyed data for each enumeration area were stored on one diskette. No more than 600 diskettes were in use at one time since

they were recycled after data verification. After the data were completely verified, all the EAs for a province (average 200 EAs per province) were transferred onto one or more Bernoulli cartridges. Approximately 10 million records (population of 9 million and 1 million households) of 52 characters in length each were saved. Sixty Bernoulli cartridges of 10 megabytes each were used to store one copy of the country's data. Five versions of the country data were kept.

Table 1. Partial list of IMPS users for population censuses

Country name	Census date	Estimated population (millions)	Operational control	Data entry	Editing	Tabulation
Benin	1989	4.3	manual	CENTRY	CONCOR	CENTS
Burkina Faso	12/85	9	part autom.	RODE/PC	CONCOR	CENTS
Burundi	8/90	5	undecided	CENTRY	CONCOR	CENTS
Cameroon	4/87	10.3	manual	IBM S/36	CONCOR	CENTS
Central African Republic	10/88	2.7	dCONTROL	CENTRY	CONCOR	CENTS
Comoros	1990	.5	manual	undecided	CONCOR	CENTS
Côte d'Ivoire	3/88	10	manual	RODE/PC	CONCOR[1]	CENTS[1]
Honduras	5/88	4.8	dCONTROL	KeyEntry III	CONCOR[2]	CENTS[2]
Malawi	9/87	7.4	manual	ICL mini	CONCOR	CENTS
Mali	4/87	8.4	manual	RODE/PC	CONCOR	CENTS
Mauritania	10/87	1.9	manual	UNKNOWN	CONCOR	CENTS
Micronesia (Pohnpei)	9/85	.30	manual	Entrypoint	CONCOR	CENTS
Niger	5/88	7.5	dCONTROL	RODE/PC	CONCOR	CENTS
Philippines	1990	57	undecided	CENTRY	CONCOR[2]	CENTS[2]
Senegal	5/88	7.5	dCONTROL	RODE/PC	CONCOR	CENTS
Somalia	11/86	7.8	manual	custom pgm	CONCOR[1]	CENTS[1]
Swaziland	8/86	.7	manual	RODE/PC	CONCOR	CENTS
Tanzania	8/88	23.5	manual	ICL mini	CONCOR[2]	CENTS[2]
U.S. (Pacific outlying areas)	4/90	.3	manual	CENTRY	CONCOR	CENTS
Yemen A.R.	2/86	9	part autom.	custom pgm	CONCOR	CENTS

[1] Program development done on microcomputers; production processing on mainframe computers.

[2] Program development done on microcomputers; have not decided what to use for production processing.

The data stored include the raw data at enumeration area level (RODE/PC file format and ASCII format), the consolidated data at province level, and two copies of the edited data.

The management of about 600 diskettes and 240 Bernoulli cartridges was accomplished effectively by using a simple external labeling system. Each diskette was clearly identified by its external label indicating an enumeration area. Equally, every Bernoulli cartridge was identified by an external label indicating a province. The task of managing and processing the census data was manageable but inconvenient, particularly because the Bernoulli Boxes were capable of holding only 10 megabytes of data at a time, and the IBM PC/AT microcomputers had only 20 megabytes of hard disk storage. Due to technical difficulties, the tape units could only be used for archival purposes and not for processing the census data.

CONCOR and CENTS do not require the entire population census file to be processed or sorted as a unit. The data can be edited batch-by-batch using CONCOR. The batches can vary in size. The edited (or cleaned) batches of data can be merged into larger batches (or geographic units) for tabulation using CENTS. The resulting cross-tabulations can be saved as much smaller files, and then consolidated to produce tables for larger geographic entities. Thus, in a system using more than one microcomputer, each machine can be processing a different geographic area simultaneously, shortening the overall processing time.

In the case of Burkina Faso, both editing and tabulation were done at the province level. The CONCOR program took an average of 20 minutes to edit each of the 30 provinces. The execution of the CENTS tabulation programs for each province took about 15 minutes. Through CENTS, the intermediate province-level tabulations were merged to produce the country-level tabulations. Three IBM PC/AT microcomputers with 20 megabytes of storage were used for editing and tabulation.

The keying of the country data in nine months instead of the estimated one year was due primarily to the high keying rates attained. The keystrokes ranged from 9 000 to 14 000 per hour with an error rate of 0.3 %. Since INSD was not able to give monetary incentives to the keyers, they provided them with other types of motivation. For example, the best keyers could choose to work during the shift (morning or evening) that was more convenient to them and schedule their leave more freely. They also were promised a permanent job at the INSD after the census was over. Given the high unemployment rate in Burkina Faso, most keyers were motivated just to keep a relatively well paying job. Out of 500 applicants only 50 were selected after taking a written test and being interviewed. Above all, the keyers were hard working and took pride in their work.

Maintenance of the microcomputers was accomplished by the data processing team. A supply of spare parts allowed them to keep downtime to only 4 %. The diskette units and the keyboards caused most of the problems. In spite of these and some initial electrical problems, the equipment maintenance was done in a timely and nondisruptive fashion. As a pioneer, INSD demonstrated that processing a population census of 9 million persons using microcomputers is not only feasible and cost-effective, but an effective approach to the production of timely results.

3.2. Senegal

The Direction de la Statistique (DS) took the second national population census of the Republic of Senegal in May 1988. The estimated population was 7.5 million.

The DS chose to purchase microcomputers for the data entry operation instead of renting IBM 3742 equipment because of the reasonable cost of microcomputer hardware, the availability of adequate data entry software, and the potential use of microcomputers for other pro-

jects after the keying operations are finished. The editing and tabulation of the data were scheduled to be done using an IBM 370/145 at the Computer Center of the Ministry of Finance, a data production facility used by other government agencies.

The original microcomputer configuration included 18 IBM PC microcomputers with 256 kilobytes of memory for data entry, two IBM PC/XT microcomputers with 10 megabytes of hard disk storage for program development, five printers, and a tape unit used for data backup. Uninterrupted Power Supply (UPS) units are used for power supply protection.

The RODE/PC software package was selected for data entry. The data would be edited using CONCOR and tabulated using CENTS 4. The CONCOR and CENTS 4 programs could be developed using the microcomputers, then transferred to the mainframe computer for production runs.

During a three-week technical assistance visit to Dakar in January 1986, data processing advisors from ISPC, under an agreement with USAID, conducted a CENTS 4 workshop, installed CONCOR and CENTS 4 on the IBM mainframe computer, and discussed tabulation plans and specifications with DS personnel.

Originally, the data processing staff responsible for the census operation consisted of the head of the data processing section and the head of the data entry section of the DS. They furthered their training during a work-study visit in early 1986 at ISPC in Washington, D.C. During a six-week stay, the data entry supervisor, who had no programming experience, learned to use menu-driven RODE/PC and assisted in developing the data entry application for the census questionnaire. The data processing chief, an experienced systems analyst with prior CONCOR and CENTS 4 training, remained at ISPC for three months. In addition to learning RODE/PC, she wrote the CONCOR edit and imputation programs. She also completed much of the tabulation programming using CENTS 4. All the work was done on an IBM PC/XT.

Upon their return to Senegal, they were able to implement modifications to the programs easily when significant changes were made to portions of the questionnaire. The data for the pilot census, conducted in March 1987, were keyed using microcomputers. The CONCOR and CENTS 4 programs were transferred to the mainframe computer. No changes to the code were required. The keyed pilot census data also were transferred from 5 1/4-inch diskettes to 8-inch diskettes to facilitate transfer to the mainframe computer.

Although they encountered no major problems while processing the pilot census data on the mainframe computer, the data processing staff became frustrated due to the inconvenience. They had to physically transport themselves and the data to the mainframe computer site several miles away. After conducting a test, they found that because of the need to share the computer facilities, it took longer to run CONCOR and CENTS 4 programs on the mainframe computer than it did on the dedicated microcomputers. The data processing staff also found the interaction with the IBM PC much friendlier than with the IBM mainframe computer.

As a result of their increasing success with microcomputers, the data processing staff decided to process the data for the full census using microcomputers instead of using mainframe computers. To upgrade the current microcomputer configuration to allow in-house processing for the full census, the data processing staff acquired extended mass storage consisting of three 20-megabyte Bernoulli Boxes and a tape drive unit. Two of these Bernoulli Boxes contain a hard disk with 80 megabytes of memory.

Although the file management of diskettes, Bernoulli cartridges, and tapes will be a critical component of the processing, the data processing staff have judged the file management

problems less formidable than the inconveniences of working in the mainframe computer environment. Additionally, the head of the data entry section feels confident about the handling of census data files because he developed and implemented the magnetic media management system used for the previous population census.

The DS acquired six additional IBM PC/AT microcomputers. Four are being used by the data processing staff for program development and for production processing. The other two microcomputers will be used by the demographers and statisticians for the analysis of the census data. DS plans to use the MCPDA for the demographic analysis of the census data.

The head of the data processing section has transferred her microcomputer expertise to three system analysts who also are working on the processing of the census data. Additionally, ISPC has been providing technical assistance to the DS in all aspects of the census processing activities. One of these areas is the monitoring of census processing activities and the production of management reports, including the preliminary counts report. The DS staff adapted the dCONTROL prototype to meet Senegal's specific needs. The preliminary counts reports which are based on manual counts of the May–June 1988 enumeration were available by October 1988. The data base and dCONTROL programs occupied less than two megabytes.

The DS plans to process a 10 % sample of the data first. However, the availability of a permanent site for storing the questionnaires and for coding and keying the data has delayed the start of the processing of the census data. The RODE/PC data entry program and the CONCOR editing programs have been tested. The CENTS tabulation programs are being finalized. A system to monitor operator performance based on operator statistics such as keystrokes per hour and keying error rates has been developed. This dBASE III PLUS system produces reports with operator statistics based on the information found in the ASCII file produced by the RODE/PC data entry program.

In the meantime, the DS staff have been using RODE/PC and CENTS for other surveys. The DS staff's enthusiasm for microcomputer use extends even further. As part of the analysis and publication stage, the DS plans to use microcomputers for thematic mapping. This is the computerized creation of maps that reflect regional variations in characteristics of the population. It requires mapping software and peripheral equipment such as a digitizer and a plotter. The U.S. Bureau of the Census will provide training.

Over the last three years, the DS has become more confident and enthusiastic about microcomputers. Not only are they using microcomputers to process the census data, but they also are using them for many other projects and surveys. Most systems analysts have microcomputers on their desks and are providing microcomputer training to junior programmers. Demographers and statisticians also are using microcomputers for their work.

With the subject-matter expertise gained during the last census, the growing microcomputer literacy, and staff enthusiasm, the prognosis for the census processing is good.

4. General Prognosis

The experiences of Burkina Faso and Senegal are representative of the positive effect microcomputers and IMPS are having on the census processing activities of many countries. Three years ago these countries were pioneers in the use of microcomputers for censuses. Now over 18 countries are using or plan to use microcomputers for some aspect of their census processing. This number is increasing rapidly as statistical offices and donor agencies realize the advantages of microcomputers.

Operational control and file management are still a challenge. As recording densities of mass

storage devices continue to increase, operational control and file management will be easier, perhaps even easier than on mainframe computers. The dCONTROL census management and reporting system is helping to resolve this challenge.

In spite of improvements, data capture is still the most time-consuming operation in census data processing. There is a limit to the rate at which data can be keyed, regardless of the speed of the software. Although one could add work stations to diminish the total time for keying, there is a point at which the logistics of such an operation become highly cumbersome. As optical mark reading (OMR) technology advances, it should become a more cost-effective data entry alternative for developing countries within the next five years.

The availability of more cost-effective tools for processing census data should make it possible for statistical offices in developing countries to focus their attention and resources on the use and dissemination of census data. Many tools are becoming available to accomplish this objective, including software packages for the production of thematic maps and desktop publishing. Additionally, hardware improvements in optical disk technology allow for easier access and transfer of large volumes of data.

Census data users themselves have more tools available which should, in principle, help relieve statistical offices of some of the responsibility they have as the principal disseminators of census data. Since many census data users have never been able to obtain census results or have obtained them five or more years after enumeration, rendering their usefulness to almost null, the availability of any data is a big improvement. Census data users will welcome the ability to access a subset of the census data on diskettes that they can use on their own microcomputer systems. It is hoped that statistical offices and census data users will focus on the analysis and dissemination of reliable data and not be distracted by the many presentation alternatives provided by the new technology.

As computer technology advances, data processing needs and software are changing. IMPS is the result of years of experience by ISPC staff, but is also the outcome of requests by census data processing managers in developing countries. As resources allow, ISPC will continue to support and modify the various modules of IMPS by critically observing its use by statistical offices in developing countries, listening to IMPS users regarding suggestions for improvements, and by taking full advantage of the microcomputer environment available in statistical offices in developing countries.

ISPC's primary objective is to make these packages usable by persons who are not highly skilled or experienced computer programmers. ISPC will make modifications to CONCOR and CENTS in the area of user-friendliness over the next few years. One goal is to make at least the definitional aspects of these packages interactive so that users need not learn a procedural language in order to use them. Some aspects of the packages, such as the definition of complex consistency edits are best stated through a procedural language. Others, such as the definition of table formats, are best done interactively. ISPC also will develop tutorials and extensive census examples to facilitate the learning process.

Enthusiasm for new technology should not diminish the effect that planning, politics, and communication have on the processing of a census. After all, microcomputers are only tools for people to use; they are not a substitute for careful planning and close coordination.

5. Conclusions

The experiences of developing countries like Burkina Faso and Senegal have made the use of microcomputers to process censuses of 10 million persons a reality. Technological developments over the last three years are making

some of the problems normally associated with large-scale processing quickly disappear. As more cost-effective tools become available for processing census data, statistical offices in developing countries should be able to focus their attention and resources on the use and dissemination of census data. We eagerly await the effect of future technological development on large-scale data processing.

6. Bibliography

Anderson, P. and Ondra, T. (1987): Trip and Assessment Report to Pohnpei, Federated States of Micronesia. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Chamberlain, K. (1988): Trip Reports to Mogadishu, Somalia. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Chamberlain, K. (1988): Trip Reports to Dar Es Salaam, Tanzania. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Gomez, E. (1988): Trip Reports to Tegucigalpa, Honduras. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Gomez, E. Cuevas, M., and Peterson, L. (1987): Trip Reports to Tegucigalpa, Honduras. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Hie, J.-M. (1988): Trip Report to Ouagadougou, Burkina Faso. Economic Commission for Africa. Unpublished report, United Nations, New York.

Le, T. and Toro, V. (1988): Trip Report to Dakar, Senegal. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Le, T. (1988): Trip Reports to Bujumbura, Burundi. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Megill, D. (1988): Trip Report to Niamey, Niger. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Megill, D. and Toro, V. (1988): Trip Report to Niamey, Niger. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Melaney, T. and Toro, V. (1988): Trip Report to Dakar, Senegal. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Rowland, S. and Melaney, T. (1988): Trip Report to Bujumbura, Burundi. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Sawaya, S. and Banerjee, K. (1987): Trip Report to Abidjan, Côte d'Ivoire. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Stroot, M. (1988): Trip Report to Sana'a, Yemen Arab Republic. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Toro, V., Le, T., and Megill, D. (1987): Trip Report to Niamey, Niger. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Toro, V., Stanecki, K., and Le, T. (1987): Trip Report to Dakar, Senegal. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Toro, V. and Melaney, T. (1987): The Use of Microcomputers for Census Processing in Developing Countries. Paper presented at the American Statistical Association Meetings, San Francisco, California. Unpublished.

U.S. Bureau of the Census (1988): Population Census Support Status of Sub-Saharan Africa. Unpublished report, U.S. Bureau of the Census, Washington D.C.

Received November 1987
Revised March 1989

Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 83–87

On a Gerardi Alternative for the Geary-Khamis Measurement of International Purchasing Powers and Real Product

Salem H. Khamis[1] and D. S. Prasada Rao[2]

Abstract: This paper briefly examines an aggregation method due to Gerardi which was used by EUROSTAT in its international comparison exercises (EUROSTAT (1977)). The Gerardi method is based on a simple geometric mean of the national prices expressed in different currency units. This is justified by the claim that the final comparisons are not affected by whether the national prices are converted or not converted before averaging. In this paper, we establish algebraically that such a claim is not valid as it involves the cancellation of zero coefficients in the numerator and denominator of a certain fraction.

Key words: National prices; purchasing powers of currencies; Geary-Khamis method; Gerardi method; solution to linear equation systems; international comparisons.

1. Introduction

The Geary-Khamis (GK) method of aggregation (Geary (1958); Khamis (1970, 1972)) is now used by the United Nations International Comparisons Project (ICP), the Statistical Office of the European Communities (EUROSTAT), the Organisation for Economic Co-operation and Development (OECD), and the regional commissions of the United Nations for inter-country comparisons of purchasing powers and real product (Kravis, Heston, and Summers (1982); EUROSTAT (1983); and OECD (1982)). The Food and Agriculture Organisation of the United Nations (FAO) also applied the GK method in calculating regional and world indexes of food and agricultural production starting in 1986 (FAO (1986)). However, the first EUROSTAT comparison of the Common Market countries for the year 1975 is based on a method due to D. Gerardi (Gerardi (1974); EUROSTAT (1977)) which is basically an unweighted version of the GK method using

[1] 23 Hillfield Road, Hemel Hempstead, Herts, U.K. HP2 4AA.
[2] University of New England, Armidale, Australia.

Acknowledgement: The authors acknowledge with thanks some useful observations by the referee on an earlier version of this paper.

the simple geometric mean of national prices to define average prices for any commodity. The GK method defines the average prices P_i of N commodities for M countries and the corresponding exchange rates R_j through the system of $M + N$ linear homogeneous equations

$$P_i^{GK} = \frac{\sum_{j=1}^{M} R_j p_{ij} q_{ij}}{\sum_{j=1}^{M} q_{ij}}, \quad i = 1, 2, \ldots, N \tag{1}$$

$$R_j = \frac{\sum_{i=1}^{N} P_i^{GK} q_{ij}}{\sum_{i=1}^{N} p_{ij} q_{ij}}, \quad j = 1, 2, \ldots, M \tag{2}$$

where p_{ij} and q_{ij} are the price and quantity of commodity i for country j.

In general the system of equations (1) and (2) has a unique positive solution for the P_i and R_j apart from an undetermined scalar multiplicative parameter. The Gerardi method used for the 1975 EUROSTAT comparison replaces the weighted arithmetic means in equation (1) by the simple geometric mean

$$P_i^{G} = \left(\prod_{j=1}^{M} p_{ij} \right)^{1/M}, \quad i = 1, 2, \ldots, N \tag{3}$$

with the corresponding definition of R_j^G similar to the GK equations (2), i.e.,

$$R_j^{G} = \frac{\sum_{i=1}^{N} P_i^{G} q_{ij}}{\sum_{i=1}^{N} p_{ij} q_{ij}}, \quad j = 1, 2, \ldots, M. \tag{4}$$

Objections to the use of a simple geometric average of national prices expressed in different national currencies were dismissed by the claim that had these prices first been converted to a common currency, the comparisons of real product and purchasing powers would not have been affected. Had the prices been converted before averaging, then the Gerardi equations (3) and (4) would have to be replaced by what we call the Gerardi-type equation system,

$$P_i^{G_1} = \left[\prod_{j=1}^{M} R_j^{G_1} p_{ij} \right]^{1/M}, \quad i = 1, 2, \ldots, N \tag{5}$$

$$R_j^{G_1} = \frac{\sum_{i=1}^{N} P_i^{G_1} q_{ij}}{\sum_{i=1}^{N} p_{ij} q_{ij}}, \quad j = 1, 2, \ldots, M. \tag{6}$$

Equation (5) is the same as equation (5.6) in EUROSTAT (1982, p. 51) and equation (6) is the same as equations (2) and (4) above and also conforms with the definition of purchasing power p_h^A in equation (4) in EUROSTAT (1983, p. 40). The argument that equations (5) and (6) above lead to the same relative ratios as those of equations (3) and (4) is based solely on a well-known property of the geometric means of equations (3) and (5) alone which are claimed to allow the cancellation of the factor $\left[\prod_{j=1}^{M} R_j^{G_1} \right]^{1/M}$ in the numerators and denominators of the ratios $P_i^{G_1}/P_s^{G_1}$ and $R_j^{G_1}/R_k^{G_1}$, thus leading to the same values as the ratios P_i^G/P_s^G and R_j^G/R_k^G obtained from the system of equations (3) and (4). This argument appears to have been accepted as a justification for not converting the national prices to a common currency before averaging them (without first showing that the cancelled factors are different from zero) as illustrated, for example, in EUROSTAT (1982, p. 51).

The main purpose of this note is to show the fallacy in the justification of the use of a simple geometric mean of national prices without conversion to a uniform currency. This fallacy is due to the fact that the cancellation of the related product of R_j's is not valid because this factor is equal to zero. In other words, it is shown below that the system of equations (5) and (6) has only the trivial solution $P_i^{G_1} = R_j^{G_1} = 0$ for all i and j.

2. Solution of the Gerardi-type Equation System

We consider the Gerardi-type equations (5) and (6) and for simplicity we drop in this section the superscript G_1 from the equations. We first observe that these equations have the trivial solution $P_i = R_j = 0$ for all i and j. We show now that, generally, this is the only solution to this set of equations. For this purpose we substitute for P_i in equation (6) its value from equation (5) to obtain, after simple algebra, for each j

$$R_j = \sum_{i=1}^{N} \frac{1}{p_{ij}} \left(\prod_{t=1}^{M} R_t p_{it} \right)^{1/M} w_{ij}, \tag{7}$$

where

$$w_{ij} = \frac{p_{ij} q_{ij}}{\sum_{i=1}^{N} p_{ij} q_{ij}}. \tag{8}$$

Assuming $R_j \neq 0$ for all j, we may divide both sides of equation (7) by R_j and, after simple manipulation, obtain

$$\left(\prod_{t=1}^{M} \frac{R_t}{R_j} \right)^{1/M} A_j = 1, \quad j = 1, 2, \ldots, M \tag{9}$$

where

$$A_j = \sum_{i=1}^{N} \left(\prod_{t=1}^{M} \frac{p_{it}}{p_{ij}} \right)^{1/M} w_{ij}.$$

The value of A_j, being independent of the exchange rates R_t, can be directly calculated for all j from the national price and quantity data. Taking logarithms of both sides of equations (9) we obtain

$$\frac{1}{M} \sum_{t=1}^{M} (\log R_t - \log R_j) + \log A_j = 0,$$
$$j = 1, 2, \ldots, M \tag{10}$$

which in matrix form is equivalent to

$$Ax = b \tag{11}$$

where

$$A = \begin{bmatrix} -\frac{(M-1)}{M} & \frac{1}{M} & \cdots & \frac{1}{M} \\ \frac{1}{M} & -\frac{(M-1)}{M} & \cdots & \frac{1}{M} \\ \vdots & \vdots & & \vdots \\ \frac{1}{M} & \frac{1}{M} & \cdots & -\frac{(M-1)}{M} \end{bmatrix} \tag{12}$$

is a square matrix of order M. Furthermore,

$$X = (\log R_1, \log R_2, \ldots, \log R_M)^T \tag{13}$$

and

$$b = (-\log A_1, -\log A_2, \ldots, -\log A_M)^T \tag{14}$$

are two column vectors of order M each, and T denotes transpose. Since the sum of each of the columns in A is zero, the rank of A is at most $M-1$. The determinant of the submatrix obtained by deleting the first row and first column has a strictly dominant diagonal and hence the rank of the matrix A is exactly $M-1$.

Accordingly the matrix equation (11) will have a solution if the rank of the augmented matrix $(A\ b)$ is equal to that of A, where $(A\ b)$

is

$$\begin{bmatrix} -\frac{(M-1)}{M} & \frac{1}{M} & \cdots & \frac{1}{M} & -\log A_1 \\ \frac{1}{M} & -\frac{(M-1)}{M} & \cdots & \frac{1}{M} & -\log A_2 \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{1}{M} & \frac{1}{M} & & -\frac{(M-1)}{M} & -\log A_M \end{bmatrix}.$$

This is an $M \times (M+1)$ matrix, whose rank is equal to that of A (i.e., $M-1$) if and only if $\sum_{j=1}^{M} \log A_j$ is equal to zero, which is generally not satisfied. This result can be proved by observing that the rank of $(A\ b)$ will be equal to $(M-1)$ if and only if there exists a non-trivial linear combination of the rows of $(A\ b)$ resulting in a zero row vector, and that the only linear combination with such a property is simply the sum, or a multiple of the sum, of the rows of the matrix.

Accordingly, the equations (5) and (6) are generally inconsistent and have only the trivial solution $R_j = P_i = 0$ for all i and j. A numerical example illustrating this result is provided in the Appendix.

3. Conclusion

The non-existence of a solution to the Gerardi-type equations (5) and (6) other than the trivial solution invalidates the argument of its being equivalent to the system of equations (3) and (4) as this argument involves the cancellation of zero factors from the numerators and denominators of ratios of the form R_j/R_k or P_i/P_s. Some objections to the use of Gerardi's equation system (3) and (4) are given in Kravis, Heston, and Summers (1982, pp. 78–79) and EUROSTAT (1982, pp. 59–61). We are not concerned here with the advantages or disadvantages of any particular system for international comparisons. We should also point out that the Gerardi System of equations (3) and (4) does lead to comparisons of real product and purchasing powers comparable numerically with those obtained by other systems of index numbers. All we assert here is that justification of averaging national prices before expressing them in terms of a uniform currency is still lacking.

4. References

EUROSTAT (1977): Comparison of Real Values of the Aggregates of ESA, 1975. Statistical Office of the European Communities, Luxembourg.

EUROSTAT (1982): Multilateral Measurement of Purchasing Power and Real GDP. Statistical Office of the European Communities, Luxembourg.

EUROSTAT (1983): Comparison of Real Values of the Aggregates of ESA, 1980. Statistical Office of the European Communities, Luxembourg.

FAO (1986): Inter-Country Comparisons of Agricultural Production Aggregates. Economic and Social Development Paper No. 61, Rome.

Geary, R. C. (1958): A Note on the Comparison of Exchange Rates and Purchasing Power Parities Between Countries. Journal of the Royal Statistical Society, Series A, Vol. 121, pp. 97–99.

Gerardi, D. (1974): Sul Problema della Comparazione dei Potesi d'Acquisto della Valute. Istituto di Statistica dell'Universita di Padova. Series papers.

Khamis, S. H. (1970): Properties and Conditions for the Existence of a New Type of Index Numbers. Sankhyā, Series B, Vol. 32, pp. 81–98.

Khamis, S. H. (1972): A New System of Index Numbers for National and International Purposes. Journal of the Royal Statistical Society, Series A, Vol. 135, pp. 96–121.

Kravis, I. B., Heston, A., and Summers, R. (1982): World Product and Income – International Comparisons of Real Gross Product. Johns Hopkins University Press.

OECD (1982): National Accounts 1951–80. Vol. 1, OECD, Paris.

Received April 1987
Revised March 1989

Appendix

A Numerical Illustration

We consider a simple example with three countries ($M = 3$) and two commodities ($N = 2$). The numbers used in this example are fictitious and the example is intended purely as an illustration of the main conclusion of the paper. The price and quantity data are given below:

Commodity i	Country j					
	1		2		3	
	Price	Quantity	Price	Quantity	Price	Quantity
1	1	10	3	4	5	2
2	2	5	4	6	6	5

The associated value share matrix is:

$$W = (w_{ik}) = \begin{bmatrix} 0.5 & 0.33 & 0.25 \\ 0.5 & 0.67 & 0.75 \end{bmatrix}$$

Following equations (9), A_k is given by

$$A_k = \sum_{i=1}^{N} \left[\prod_{j=1}^{M} \left(\frac{p_{ij}}{p_{ik}} \right)^{1/M} \right] w_{ik}, \quad k = 1, 2 \text{ and } 3.$$

Numerical values of A_1, A_2 and A_3, for the price-quantity data above, are

$$A_1 = 2.123 \quad A_2 = 0.8812 \quad A_3 = 0.5808.$$

Therefore,

$$A_1 A_2 A_3 = 1.0866$$

and

$$\log A_1 + \log A_2 + \log A_3 = 0.083.$$

In this case the condition that $\sum_k \log A_k$ should be equal to zero, necessary for the existence of a nontrivial solution to the Gerardi-type equations (5) and (6), is violated. In general most data sets would result in non-zero values for $\sum_k \log A_k$ unless the expenditure ratios, w_{ik}, are identical across all the countries.

Journal of Official Statistics
Vol. 5. No. 1, 1989, p. 89

Letters to the Editor

Letters to the Editor will be confined to discussions of papers which have appeared in the Journal of Official Statistics and of important issues facing the statistical community.

Comments on Cohen, Xanthopoulos, and Jones

The article by Cohen, Xanthopoulos, and Jones (JOS, Vol. 4, No. 1) contains what I consider to be an unfortunate error. The authors maintain that regression analysis in survey work assumes the usual form of the multiple regression model with independent and identically distributed errors (pp. 19, 20). It does not. If it did, then one could use OLS without fear.

The usual design-based approach is to avoid introducing the regression model all together and to treat the full population regression coefficient, $(X'X)^{-1}X'Y$, as the goal of estimation (for example, see Shah, Holt, and Folsom (1977)).

An alternative model-based approach might invoke the usual multiple regression model but allow for the possibility of an error structure with a complicated pattern of correlations within primary sampling units (PSU's) and maybe even across PSU's within strata. In this model-based framework, the inclusion of weights in the regression estimator is driven by the fear that the model may be missing regressors (Holt, Smith, and Winter (1980)).

References

Cohen, S. B., Xanthopoulos, J. A., and Jones G. K. (1988): An Evaluation of Statistical Software Procedures Appropriate for the Regression Analysis of Complex Survey Data. Journal of Official Statistics, 4, pp. 17–34.

Holt, D., Smith, T. M. F., and Winter, P. D. (1980): Regression Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Series A, 143, pp. 474–487.

Shah, B. V., Holt, M. M., and Folsom, R. E. (1977): Inference About Regression Models from Sample Survey Data. Bulletin of the International Statistical Institute, 47, pp. 43–57.

Phillip S. Kott
National Agricultural Statistics Service
U.S. Department of Agriculture
Washington, D.C.
U.S.A.

Reply

Our intent was to state the classical assumptions and to indicate how complex survey designs depart from them. The paper then moved on to the evaluation of statistical software procedures appropriate for the regression analysis of complex survey data.

Steven B. Cohen, Ph. D.
Senior Research Manager
National Center for Health Services Research and Health Care Technology Assessment
Rockville, MD
U.S.A.

Journal of Official Statistics
Vol. 5. No. 1, 1989, p. 91

Special Notes

Nominations Invited for Distinguished Statistical Ecologist Award

The Awards Committee solicits nominations for the Distinguished Statistical Ecologist Award of the International Association for Ecology/Statistical Ecology Section. All members of the statistical community are encouraged to submit nominations. Letters of nomination should indicate why the nominee is especially deserving; additional documentation (e.g., curriculum vitae) may be included when appropriate. The nominee may be from any country and need not be a member of INTECOL.

The members of the Award Committee are: L. Orloci (Canada), O. Rossi (Italy), G. P. Patil (U.S.A., Chair), and D. Simberloff (U.S.A.). The first recipient of the award was G. P. Patil, who was honored at the Fourth International Congress of Ecology in 1986 at the time of his Plenary Lecture on Statistical Ecology to the Congress.

The recipient(s) will be invited to participate in the Fifth International Congress of Ecology, August 23–30, 1990, Yokohama, Japan.

Send nominations, no later than September 15, to:

Professor G. P. Patil
Chair, INTECOL
Distinguished Statistical Ecologist Award Committee
Center for Statistical Ecology and Environmental Statistics
Pennsylvania State University
University Park, PA 16802, U.S.A.

Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 93–104

Book Reviews

Books for review are to be sent to the Book Review Editor Jan Wretman, Statistical Research Unit, Statistics Sweden, S-115 81 Stockholm, Sweden.

Wainer, H. (Ed.), Drawing Inferences from Self-Selected Samples. Springer-Verlag, New York, 1986. ISBN 0-387-96379-0 (Springer-Verlag New York), ISBN 3-540-96379-0 (Springer-Verlag Berlin). xii + 163 pp., DM 45.00.

Most statisticians analyze data through models that describe an underlying population of interest, for example the iid normal model:

$$y_i \sim_{\text{iid}} N(\mu, \sigma^2). \quad (1)$$

In practice data come in the form of samples. Let $s_i = 1$ if unit i is selected and 0 otherwise. Then we actually see the distribution of Y given that $S = 1$:

$$y_i | s_i = 1 \sim ? \quad (2)$$

Statistical texts usually ignore the implicit conditioning on S and replace ? in (2) by the population model, such as (1). This is fine if Y and S are independent, a reasonable assumption if we have (or plausibly can pretend we have) a simple random sample. However often this is not reasonable. An important case is *self-selection*, when the subject's choice enters into the inclusion process. *Drawing Inferences from Self-Selected Samples* (DISS) presents analytical approaches to self-selection in four application areas: (i) comparisons of SAT scores when scores are available only to those who choose to take the test (by Howard Wainer); (ii) evaluation of methadone clinics in the treatment of heroin, where clinic patients are volunteers (by Burton Singer); (iii) assessing the effects of job training, where data are available on those who choose to train (by James Heckman and Richard Robb); and (iv) survey nonresponse, where the sample is restricted to individuals who choose to respond (by Robert Glynn, Nan Laird, and Donald Rubin). The book also includes lively contributions from two distinguished discussants, John Hartigan and John Tukey.

I would like to have seen the views of a survey sampling specialist (one who, in my trivial example, treats Y as fixed and bases inference on the distribution of S given Y). The sampler is trained to handle problems of prob-

ability sampling where the distribution of S is under control of the sampler but may not correspond to simple random sampling. Sampling theory seems to me of limited help in self-selection, where the distribution of S is not under our control. Nevertheless I suspect a sampler would provide an interesting counterpoint to some of the viewpoints in the book, for example, the remarks on panel vs cross-sectional survey designs in Heckman and Robb's paper. On a point of detail, samplers might also take issue with Heckman and Robb's statement that many social science data sets contain hundreds and thousands of *independent* observations (page 67, italics mine), since large samples usually involve clustering that leads to correlated observations.

The chapters by Wainer on SAT scores and Singer on methadone treatment assessment are well written and relatively nontechnical, although some of the concepts discussed are subtle. They provide a valuable introduction to the more technical material in the papers that follow. Wainer compares mean SAT scores for 21 U.S. states that give primarily SAT tests ("SAT states") with mean SAT scores for 29 states that administer primarily the American College Testing (ACT) Program ("ACT" states). Scores for the latter group are much higher, the distributions barely overlapping. Does the difference (a) reflect superior SAT-taking ability in the ACT states, or (b) are SAT-takers in the ACT states a more select group? Wainer shows that SAT-takers in ACT states rank higher in their college class than SAT-takers in SAT states, thus providing strong evidence in favor of (b).

A common statistical approach to lack of comparability between treatment groups is covariate adjustment on variables thought to capture (or at least reduce) the lack of comparability. Wainer discusses this approach to his problem, using the covariate "percent of high school seniors taking the SAT," or participation rate for short. He points out this approach may yield bias, since "higher-quality schools will yield a greater proportion of their SAT-taking pool," that is, participation rate is a measure of SAT-taking ability as well as of the selection effect. In path analysis terminology, participation rate is not causally prior to SAT-taking ability. Wainer applies an alternative strategy based on converting ACT scores to SAT scores using equating information. This approach almost by magic removes the differences in scores in the SAT and ACT state groups: the adjusted medians are identical! However equating is a not a viable option in other settings. Econometricians would perhaps advocate attempting to fix covariate adjustment via structural equation modeling. It would be interesting to see how results from such an approach compare with the answers from equating, which might plausibly be treated as the gold standard here.

Singer's chapter on Methadone Maintenance Treatment (MMT) programs first reviews heroin abuse and treatment in Hong Kong, Sweden, and New York and presents a pattern of heroin addiction which serves as a baseline for comparing treatments. Singer then discusses some intervention strategies, and the evaluation of MMT programs via performance-based ratings indices. With regard to self-selection, Singer raises the important question of whether the aim is to evaluate and compare programs on the population of volunteers who enter them, or on the target population of all addicts, an issue which is also stressed in the Heckman and Robb paper. Although he views inference to all addicts as desirable, he judges it impractical given the lack of quantitative knowledge to distinguish the two populations. Even if such knowledge were available, it seems to me that considerable extrapolation would be involved.

Even the problem of comparing clinics on the volunteer population is far from easy, given that different clinics may attract different clients. Singer's approach is to develop relatively crude indices based on changes in a patient's activities and behavior before and after treatment, and in essence to compare these indices with historical control treatments deemed to have been successful. Singer discusses the limitations of this approach. The use of change measures from longitudinal data, allowing each subject to act as his or her own control, seems to me a time-honored and powerful way of alleviating the effects of self-selection. In addition, some form of simple covariate adjustment on the baseline variables would seem to be feasible here if they are consistently measured across studies, but maybe that is a big if.

The paper by Heckman and Robb provides an extensive model-based analysis of the selection bias problem in the context of the impact

of job training programs. It is not in fact the one presented at the conference, which was published elsewhere, but a revision that takes into account the skeptical reactions of Tukey and Hartigan to the conference paper. (Readers might be able to estimate a Hartigan/Tukey effect by constructing a measure of change from the two versions!) The Hartigan and Tukey discussions are published before the revised Heckman-Robb paper, preserving the chronological order. (Since we all read discussions of papers before papers, maybe this is the right order anyway!) The discussions focus on the sensitivity of results to untestable assumptions in the Heckman and Robb analysis, and illustrate the differing attitudes of statisticians and econometricians towards statistical analysis. Econometricians start with a theoretical model and work towards the data, pruning parameters until the model can be estimated. Statisticians start with the data and work towards a theoretical model, estimating parameters that shed light on, but may only approximate, idealized quantities of econometric theory. Econometricians complain that the statisticians are too atheoretical, or "context-free." Statisticians complain that econometricians are too insensitive to the limitations of their models and the data. If the bridge between data and theory is shaky, as in the self-selection problem, then these two approaches can end up in different places.

Heckman and Robb's paper is long and technical, but contains a helpful introductory section that clarifies its objectives. These include (i) a careful definition of the parameter of interest, the effect of job training; (ii) specification of minimal assumptions needed to identify the parameter, for single cross-section, repeated cross-section and longitudinal designs, under conventional and enriched behavioral models of earnings. They conclude that "although longitudinal data are widely regarded in the social science and statistical communities as a panacea for selection and simultaneity problems, there is no need to use longitudinal data to identify the impact of training on earnings if conventional specifications of earnings functions are adopted. Estimators based on repeated cross-section data for unrelated persons identify the same parameter. This is true for virtually all longitudinal estimators." (Page 65).

This conclusion has created some controversy among advocates of panel surveys, since the implication is that panel surveys are overused in practice. However, the practical ramifications are not clear, since the conclusion results from a mathematical analysis that focusses on the identifiability issue in the context of specific selection models. Survey sampling arguments for panel surveys consider a separate issue, the sampling error of simple estimates of change (such as the difference in means) that ignore selection effects entirely (Cochran (1977, Section 12.10)). The Heckman and Robb analysis ignores sampling error entirely. Also a major reason for social science panel surveys (such as the U.S. Survey of Income and Program Participation) is their ability to measure micro-level transitions that are inestimable from repeated cross-sections.

Heckman and Robb's approach to selection bias is to model the data and the selection mechanism. In the context of job training, a parameter α is introduced to represent the additional earnings from training. Training occurs if an observed variable called index of net benefits (IN) crosses a threshold, say zero. In the behavioral model, IN is viewed as the difference between the expected benefits of training (the gain in future earnings, discounted to some degree) and the expected costs (expenses and loss of earnings during training). Selection bias arises under this model when the propensity to train is related to future earnings in the absence of training, after adjusting for the effects of observed covariates. For example, if (given covariates), those predisposed to be successful are more likely to train, the positive effects of training will be exaggerated by comparing the (adjusted) mean incomes of trainees and non-trainees.

Writing down formal models such as those considered by Heckman and Robb can be a useful way of clarifying thinking. However the purely economic model of training choice seems hard to swallow, as does the assumption of a constant training effect for all individuals: Heckman and Robb do provide a limited discussion of random training effects, but it is mainly directed at defining the parameter of interest. If a distribution of training effects exists, it seems to me that longitudinal data would be needed to estimate it, so the assumption of constant training effect favors the cross-sectional design.

Heckman and Robb's cross-sectional methods for adjusting for selection bias appear to depend crucially on finding instrumental variables that are predictive of the decision to train, but not predictive of earnings. No specific suggestions for variables are offered, and (coming from the statisticians' camp) I have less confidence than Heckman and Robb in the ability of econometric theory to supply them. Purely for illustration, let me propose the variable "distance to training site." This variable might be strongly related to the decision to train, and a plausible econometric story might justify the assumption that "distance to training site" is not related to earnings, after adjustment for other exogenous variables in the model. Human populations are heterogeneous, however, and social science theory does not lead to all-encompassing physical laws. Thus it also seems plausible that the variable *is* related to earnings, particularly if it is acting as a proxy for some unmeasured geographical covariate. This difference of opinion does not matter much for some types of analyses, but it does if a large selection bias adjustment rests on it. For me, these instrumental variables (IV's) often supply blood to a body that is already dead; "minimal identifying assumptions" (MIA's) are too often "missing in action"!

What are the alternatives to selection modeling? One approach is to try to collect as many variables as possible related to the selection process, and then use these variables in a standard covariate adjustment. Here longitudinal surveys may have a distinct advantage over cross-sections, because of superior ability to collect time series information.

Selection models of the type considered by Heckman and Robb have also been applied to survey nonresponse, and it is this application that is the subject of the Glynn, Laird, and Rubin chapter. These authors compare two modeling strategies; let Y denote the outcome variable of interest, X fully-observed covariates and R an indicator for response ($R = 1$) or nonresponse ($R = 0$). *Selection modeling* writes the joint distribution of Y and R in the form

$$f(R,\ Y|X,\ \theta,\ \psi) = f(Y|X,\ \theta)f(R|Y,\ X,\ \psi), \tag{3}$$

where the first component characterizes the distribution of Y given X in the population, and the second component models the incidence of nonresponse as a function of X *and* Y. *Mixture modeling* writes the joint distribution in the alternative form

$$f(R,\ Y|X,\ \xi,\ \omega) = f(Y|X,\ R,\ \xi)f(R|X,\ \omega), \tag{4}$$

where the first distribution characterizes the distribution of Y given X in respondent and nonrespondent strata, and the second component models the incidence of nonresponse as a function of X only. The distribution of Y given X is then a mixture of the distribution of Y given X in the response and nonresponse strata, which explains the name. Selection modeling is natural to econometricians since their models relating Y and X are formulated in the unrestricted population. Mixture modeling is perhaps more natural for statisticians since it is closer to the structure of the observed data. In particular the mixture modeling form (4) emphasizes a basic difficulty inherent with the data; since there are usually no data on Y for nonrespondents, there is no information for estimating the distribution of Y given X, $R = 0$. Rubin (1977) relates the distribution for nonrespondents to that for respondents using a Bayesian prior distribution. The selection modeling form (3) can be estimated without explicit inclusion of prior information relating respondents and nonrespondents. However such a prior specification is implicit, and sensitivity to model specification is an equally serious problem for either version of the model.

Glynn, Laird, and Rubin display sensitivity of the selection approach to model misspecification by simulating results under correctly-specified and misspecified models. They conclude that the method is very unstable unless a covariate is available that is related to *only* one of response or outcome; the variable plays the analogous role to the IV variables in the Heckman and Robb paper. Here as in the job training context, the key question is whether such variables can be found in practice: Glynn, Laird, and Rubin are pessimistic.

The chapter also compares the selection modeling and mixture modeling approaches when a subsample of nonrespondents are available via follow-ups. The assumption is made that the subsample is random. Comparisons are made using simulated data, and real data from a survey on drinking behaviors. They conclude that mixture modeling is more robust than se-

lection modeling to departures from distributional assumptions.

Like any collection of papers, DISS lacks some degree of cohesiveness. The book presents the views of distinguished applied statisticians on a problem that arises in the real world, rather than the artificially constructed world of many mathematical statistics texts. I found the book stimulating and recommend it.

References

Cochran, W. G. (1977): Sampling Techniques. 3rd Edition. John Wiley, New York.

Rubin, D. B. (1977): Formalizing Subjective Notions About the Effects of Nonrespondents in Sample Surveys. Journal of the American Statistical Association, 72, pp. 538–543.

Roderick J. A. Little
U. C. L. A.
Los Angeles, CA
U.S.A.

Turner, C. F. and Martin, E. (Eds.), Surveying Subjective Phenomena, Volume I and Volume II. Russell Sage Foundation, New York, 1984. ISBN 0-87154-882-8 (Vol. I), 0-87154-883-6 (Vol. II), 0-87154-881-X (set). xvi + 495 pp. (Vol. I), xvi + 618 pp. (Vol. II).

It is astounding how much attitude survey research has become a part of academic and political life. It is estimated that a minimum of 100 million survey interviews were conducted between 1971 and 1976 in the United States. More than 28 million survey interviews were conducted by telephone during 1980. It is estimated that 39 % of the British public have been surveyed. In a single one-month period there is documented evidence of the distribution of more than 200 million copies of poll stories in American news media and more than 50 million copies in Britain. More than half of the published articles in the field of sociology report survey data, as do about 30 % of the articles published in political science and economics.

The two volumes reviewed here represent the proceedings of a multi-year panel on "Survey Measurement of Subjective Phenomena," convened in 1980 under the auspices of the United States Committee on National Statistics. The committee was convened because of the discovery of "several instances in which seemingly equivalent (public opinion) survey measurements made at approximately the same time produced surprisingly different results (I:xiii)." The problems all clearly involved nonsampling sources of error (as opposed to sampling errors which are handled by probability theory and confidence testing). The purpose of the panel was to study "the use, reliability, and meaningfulness of survey measurements of attitudes, opinions, and other subjective phenomena (I:xiv)." The work of the panel took more than two years to complete.

In the course of its work, the panel stimulated so much interesting research on the survey profession qua profession that there may be enough material for a unit on subjective measurement in a course on the history of ideas. Volume II Chapter 1 is a fascinating historical sketch of the different kinds of attitude research that developed in the early years – social distance scales, Thurstone scales, Likert's "fast" scaling technique, and so on. A careful reader also finds in this chapter the seeds for debates between academic disciplines about which one studies "real" attitudes and why studies from perspectives other than one's own are to be criticized as "conceptually inadequate." Volume II Chapter 10 is a comparison of the tendency for particular survey houses, e.g., Gallup, Harris, etc., to prefer particular approaches to questionnaire construction, e.g., open-ended, middle response categories, etc., Volume II Chapter 2 is a similarly fascinating sketch of the attempts of economists to define "utility" in a way that is not circular and therefore incapable of independent measurement. The author concludes that "economists have been more concerned with drawing out the implications of utility assumptions based on casual introspection or on an a priori conception of rationality than with attempting to measure utility in practice (Vol. II p. 42)." This level of insight and intellectual honesty in a book not specifically attempting to discredit economic analysis is refreshing.

Developments or events in four specific areas served as catalysts for *Surveying Subjective Phenomena:* (1) Surveys of public confidence in the leaders of national institutions, done at the same time and using allegedly equivalent mea-

sures, showed substantial discrepancies in both the levels of reported confidence and the trends across time. (2) Trend studies of "happiness" indicators showed divergent results depending on the survey organization conducting the field work. (3) Surveys of public attitudes toward science were being openly criticized for reifying public opinion, i.e., putting words in the mouths of respondents – on topics for which there was little public information or understanding. (4) Specific surveys were the targets of attack by non-social science university faculty as being based on a methodology that was "ambiguous," "meaningless," and "prejudicial."

Surveying Subjective Phenomena attempts to come to grips with the issues that each of these criticisms raises for the survey measurement profession. Volume I Chapter 1 points out that the problem of fallible measurement is not limited to subjective indicators, survey research, or even social science. There is a fascinating discussion of interlaboratory experiments conducted to achieve replicated measurements of natural science physical constants. The measurements *ought* to have produced the same result but they did not. Experimental studies of the variability among mesurements made by different scientists, by different laboratories, and by different analytical procedures led to "a better understanding of the error structure of such measurements (Vol. I p. 16)." These experiments in the physical sciences are the basis for the panel's recommendation for a coordinated interlaboratory program of measurements for survey research.

In a number of ways the work is a "stiff-upper-lip" exploration of the soft underside of public opinion and survey research. The work reviews past, well-publicized "failures" of survey research, e.g., the Literary Digest poll and the Dewey-Truman polls in 1948. The work reviews, in detail, the recommendations made by blue-ribbon commissions convened to study and make recommendations about those past failures (Mosteller et al. (1948)). The work reviews, chronologically and in detail, attempts made by professional polling associations, other organizations or individuals, and even the Federal government to develop, implement, and enforce standards to ensure quality and consistency in the public opinion or survey product. It was enlightening to me to see that so many of the recommendations and findings of the panel's work have been part of previous efforts as well.

The work raises many of the same questions about nonsampling errors, e.g., intensity of opinion, manufactured responses, question wording bias, questionnaire order effects, respondent understanding, selective reporting of results, and clarity of the concept being measured – as are covered in more polemical tours of the same horizon such as *Lies, Damn Lies, and Statistics* or *The Pollsters* (Wheeler (1976) and Rogers (1954)). But *Surveying Subjective Phenomena* explores the issues more fully and, in my opinion, in a more balanced fashion.

Volume I of *Surveying Subjective Phenomena* contains the panel report per se. Chapters in this volume are compilations of sections contributed by panel members and then subjected to the critical eye of editors and outside reviewers. Some of the sections and chapters in Volume I are designed to review the research and results on various topics in the literature on the reliability and validity of subjective survey measurements. Other sections and chapters in Volume I go considerably astray from this goal and are presented as new ideas or new methodologies that the panel recommends for analyzing subjective survey measures. Finally, there is a 30-page list of panel recommendations to producers and users of subjective survey data.

Volume II consists of individual contributions of authors commissioned to undertake special studies. These chapters were also subject to outside review. Some of these chapters bear directly on the point and mission of the two-volume work – some are in-depth studies, not published elsewhere, of issues in the reliability and validity of subjective survey measures. Volume II Chapter 8, for instance, is a summary of the "non-attitudes debate." Other chapters, however, go considerably far afield from this goal.

As with any published compilation of this scope (and of this panel-based methodology), the work is excessively long; overwritten at some points, underwritten at others; and, overweighted toward the interests (or abilities) of those who happen to have been panel members. Everybody who reads the work will find something that is of great interest, but they also are likely to find a great deal that is not. The two volumes are a good first draft of a book that should be about one third of its 1 145 page length.

On the other hand, I sincerely believe that there is little that has ever been said or written about the reliability and validity of subjective survey data that does not appear somewhere in this work. For this reason, *Surveying Subjective Phenomena, Vols. I & II* rivals the scope of other excellent works with a similar mission (Rossi, Wright, and Anderson (eds.) (1983)) and, therefore, merits attention by practitioners and students of social surveys.

A number of very general observations can be made about the areas of success or failure of the book. These successes and failures, I hope, will define the shape of research on survey methodology in the coming years.

The work makes an important three-way distinction between subjective phenomena, facts, and quasi-facts: (1) Subjective phenomena are those that, in principle, can be directly known, if at all, only by persons themselves – such as expectations (to vote, to have children), satisfactions (happiness, utility), subjective judgements (confidence, fairness), or opinions (for or against something). (2) Factual measurements are in principle verifiable without reference to respondents' interpretations. (3) Quasi-factual measurements allow latitude for the respondent's definition of the criterion for the (factual) behavior or event in question – such as unemployment (whether or not one is actively looking for a job), housing quality (whether a unit is deteriorating or sound), neighborhood quality (what boundaries), crime victimization (whether or not an encounter is judged to be an assault), or ethnicity (judgements based on language, father's lineage, mother's lineage, or other factors.)

The vital issues in survey measurement in this work have to do with measurement of subjective and quasi-factual phenomena. The "true score" models of physical scientists and psychologists are beside the point when one has to consider how to design experiments and calibrate the sources of measurement error for subjective and quasi-factual phenomena. The point of the definition of subjective and quasi-factual phenomena is that there is not an externally verifiable true score. Therefore it is somewhat surprising that the introduction to measurement error in Volume I Chapter 4 is a mechanistic retread of the "true score" model. The panel loses a valuable opportunity to introduce a mathematical notation and language for error models that would contribute significantly to the literature.

Beginning with Volume I Chapter 5 and continuing through much of the rest of the work, *Surveying Subjective Phenomena* concentrates, a section at a time or a chapter at a time, on the work of specific individuals or groups of authors. Volume I Chapter 5, for instance, summarizes early work (e.g., Cantril), contemporary work (e.g., Schuman, Presser, and Associates), and new results showing empirical patterns of disagreement among subjective survey questions. The topics covered in this chapter and in the follow-up piece in Volume II Chapter 7 are not as extensive, nor the analysis as probing, as the book-length treatments of these topics that are summarized in the chapter or that have been published subsequently.

Volume I Chapter 6, on the other hand, is a previously unpublished, mathematically advanced analysis of survey data using the Rasch model for item-centered and respondent-centered analysis. One wonders why this chapter is included and the reason for its placement in Volume I of *Surveying Subjective Phenomena*. The other chapters discussing measurement error do not hint at a Rasch model solution. They are not written in a way to motivate its selection as a tool to manage the complexities of analysis and interpretation that are brought forth. The Rasch model is brought forward to "call attention to some approaches to scientific analysis of survey data that are either novel or underused (Vol. I p. 179)." This is a weak justification, and its presentation seems out of place.

A number of other chapters, notably in Volume II, have this same feeling of being out of place. The reader is struck by how very little relationship there is between the mathematically technical chapters on models of subjective error measurement and the inductive, analytic chapters on patterns of results. The technicians in the subjective error field seem to be stuck on problems that have to do with measuring item and category response metrics (Rasch, latent class, etc.). On the other hand, the inductivists (who actually design and administer a lot more surveys) are stuck on problems of conceptual clarity and definition.

A number of chapters or sections raise, for the record, significant issues in subjective measurement, but have little to say beyond acknowledgement of the problem. Volume I Chapter 7, for instance, raises the issue of con-

ceptual ambiguity in surveys: "if the concepts used in survey questions are not understood in the same way by the survey researcher and the respondent, then responses to the questions are likely to be misinterpreted by the researcher (Vol. I p. 235)." The chapter does not make a clear statement on this issue. It consists of: (1) an extremely general note on the definition of "public opinion"; (2) a lively discussion of what people mean when they use the word "risk"; (3) some examples where interpretations of questions apparently were influenced by the set of response categories; and, (4) an extremely brief description of a technique called ethnographic semantic mapping. The reader, I believe, would be better off with more discussion of this topic, or less. It is as if the editors did not know where else to put these pieces, did not want to leave them out, and were not suffering from the discipline of an overall page limitation.

Volume I Chapter 8 is a review of what is known about the effect of respondent-interviewer social dynamics during an interview. The conclusion is that "variability in the social aspects of the interview situation results in variability in respondents' role expectations and behaviors during interviews (Vol. I p. 273)." But the patterns are inconsistent: "we have only begun to understand how the interview, viewed as a social relationship, influences responses to survey questions (Vol. I p. 274)." Volume II Chapter 9 explores a related terrain: "social desirability ... the notion that some things are good and others are bad, and the notion that respondents want to appear "good" and answer questions in such a manner as to be perceived that way (Vol. II p. 258). The conclusion is similarly vague: "conceptual ambiguities plague the notion of social desirability (Vol. II p. 276)." These chapters, like many others, are notable for their examples but not for their conclusions. After reading several hundred pages like this, one begins to grasp how subjective measurement is; a field rich with rules of thumb about survey design (and other forms of folklore), but a "scientific" understanding of the process may be so complex and so expensive that it will never be achieved.

Volume I Chapter 9 consists of notes on "psychological" sources of bias in the question and answer process. Some of these sources of bias include subtle cues of grammatical structure, affective connotations of particular words, or mood changes induced by positively- or negatively-worded questions. As a sociologically-trained researcher, I find this one of the more fascinating chapters because the point of view on the nature and competence of the respondent is so different from what I am used to. One of the lines of research reviewed in the chapter takes a bald position against the use of any sort of introspective reports because "people do not necessarily have privileged knowledge of their own attitudes, motives, or the causes of their behavior (Vol. I p. 298)." In a year of mud-slinging election advertising in the United States, I find fascinating the suggestion in another study reviewed in this chapter that people often cannot verbalize the reasons for their likes and dislikes because "the salient and notable features of an object are not necessarily the same features that feelings are attached to (Vol. I p. 299)." The conclusion of the chapter attempts to strike the ball right out of the park: "uncritically accepting respondents' stated purposes or motives as valid and basing a full-scale analysis on them is a risky strategy, at best (Vol. I p. 300)." It is too bad that the entire book was not more cohesively constructed so that some of the implications of the statements made in this chapter could be explicitly addressed in other parts of the discussion.

Volume I Chapter 10 contains the 18 recommendations from the panel's multi-year effort. I will not summarize them here because they do not flow directly from the preceding 300 pages nor from the 800 pages of special studies that follow. The recommendations have mostly to do with how the profession of survey research ought to be institutionalized, managed, funded, and monitored in a market economy. Those who want recommendations on how to do surveys better will have to look at the research results in individual chapters and not at the panel's 18 recommendations.

The panel advocates the strongest possible recommendations regarding public education, industry regulation, and subsidies for methodological research. The panel hopes its recommendations will offset some facts about the survey profession: (1) public opinion polling is a competitive industry in the United States and in other countries; (2) large sums of money are not likely to be forthcoming from private sources for methodological research, and, (3)

neither national governments nor professional associations are in a position to enforce strict guidelines for the polling profession. Given these facts, it is unclear what effect the panel and its list of recommendations will have on the priorities and conduct of the survey profession.

References

Mosteller, F., Hayman, H., McCarthy, P., Marks, E., and Truman, D. (1949): The Pre-Election Polls of 1948. Social Science Research Council, New York.

Rogers, L. (1954): The Pollsters. Basic, New York.

Rossi, P. H., Wright, J. D., and Anderson, A. B. (eds.) (1983): Handbook of Survey Research. Academic Press, New York.

Wheeler, M. (1976): Lies, Damn Lies, and Statistics. Liveright, New York.

D. Garth Taylor
Chicago Urban Leage
Chicago, IL
U.S.A.

Dey, A., Orthogonal Fractional Factorial Designs. Wiley Eastern Limited, New Delhi, 1985. ISBN 0-85226-165-9. viii + 133 pp., £8.40.

Even the most advanced courses in the design of experiments do not go very deeply into fractional factorials, apart from the traditional 2^{n-k} and 3^{n-k} series. There may perhaps be a passing reference to the 4^5 and 5^6 orthogonal main effects plans based on sets of mutually orthogonal latin squares, but there is rarely time for anything more.

The few topics mentioned above represented the frontier in fractional factorials until the 1960s. Little further progress was made on asymmetrical fractions until the work of Addelman and Kempthorne (1961 a and b) and Margolin in (1968; 1969 a, b, and c; 1972). Since then considerable advances have been made by several statisticians, including Professor Dey himself. Their work has appeared in various journals, among them *Technometrics*.

The interest of *Technometrics* in this work should not be surprising because the past decade has seen a surge in the use of orthogonal fractions by engineers who are involved in modern quality assurance and process improvement. Until lately, they have had available to them only orthogonal main effects plans for two and three factors, whose derivation has often been wrongly attributed to Taguchi. But the main effects plans are not enough: engineers need to have access to good resolution IV designs.

Dey has gathered together the results on orthogonal fractional factorials obtained in the past forty years or so, producing a short but useful synthesis. There is an enormous amount of interesting information, especially about asymmetrical fractions. Reasoning, no doubt, that the interested reader can find the derivations of the procedures in the original papers, the author does not repeat the proofs. However, he does provide some examples of those techniques, including, for example, a helpful discussion of the derivation of the design of Bose and Bush for $3^9//27$ (meaning 9 factors at 3 levels in 27 runs). It is good to see all these methods brought together in one volume.

Unfortunately, retrieval of the information in the book is a problem. The author has added tables, which he calls indexes, that are intended to help the reader find in the text procedures for constructing appropriate designs. I tested them on two examples. First, I tried to find the lattice for $3^7//18$. It was not listed in Table 2.3, "Index of Orthogonal Main-Effect Plans for Symmetrical Factorials," which referred me instead to the 16 run fraction by Stark, but did not tell me where to find it. Happening to know that this lattice can also accomodate a two level factor, I next looked up the $3^7.2$ lattice with 18 points in Table 3.4, "Index of Orthogonal Main-Effect Plans for Asymmetrical Factorials," and was referred to Section 3.3, where I failed to find it. Finally, I tried $6.3^6//18$ and was sent to Section 3.4.3 where I found it on pages 58 and 59. There Dey mentions that it was derived from the lattice of Addelman and Kempthorne (1961 b), but does not tell where to find the derivation of that design. (It is actually derived in a well-written section starting on page 29). Obviously, this book sorely needs a proper index.

My second attempt was to find a resolution IV design for $3.2^4//24$. I found the reference in

Table 4.5, "Index of Orthogonal Resolution IV Designs," which referred me to Section 4.3.1. There I found a procedure for $t.2^{n-1}$ fractions, and could have used some help. After staring at it for a while it dawned upon me that I should attack it as a 6.2^3 problem. The information is there, but it is hard to find.

This is an interesting book and I am glad to have read it, but I wish it were not so condensed. The first chapter is a very short introduction to the topic, marred by several typographical errors. The author's style is so terse that any but the mathematically sophisticated reader will find it hard work indeed. This book could prove helpful to the mathematical statistician who is engaged in experimental design and might be asked sooner or later for an orthogonal resolution IV fraction of an asymmetrical factorial. I am not sure that this book will provide the answer, but it should point the right direction in the literature.

References

Addelman, S. and Kempthorne, O. (1961 a): Orthogonal Main Effect Plans. Aerospace Research Laboratories Report No. 79.

Addelman, S. and Kempthorne, O. (1961 b): Some Main Effect Plans and Orthogonal Arrays of Strength Two. Annals of Mathematical Statistics, 32, pp. 1167–1176.

Margolin, B. H. (1968): Orthogonal Main Effect 2^n 3^m Designs and Two Factor Interaction Aliasing. Technometrics, 10, pp. 559–573.

Margolin, B. H. (1969 a): Results of Factorial Designs of Resolution-IV for the 2^n and 2^n 3^m Series. Technometrics, 11, 431–444.

Margolin, B. H. (1969 b): Resolution-IV Fractional Factorial Designs. Journal of the Royal Statistical Society, Series B, 31, pp. 514–523.

Margolin, B. H. (1969 c): Orthogonal Main Effect Plans Permitting Estimation of All Two Factor Interactions for the 2^n 3^m Series of Designs. Technometrics, 11, pp. 747–762.

Margolin, B. H. (1972): Non-Orthogonal Main Effect Designs for Assymmetrical Factorial Experiments. Journal of the Royal Statistical Society, Series B, 34, pp. 431–440.

Peter W. M. John
University of Texas
Austin, TX
U.S.A.

Kapadia, R. and Andersson, G., Statistics Explained. Basic Concepts and Methods. Ellis Horwood Limited, Chichester, 1987. ISBN B-7458-0053-X, 0-7458-0315-6, 0-470-20966-6. 234 pp., £ 12.95.

In the field of statistics, the bulk of the literature discusses statistical methods and techniques and the intended audience is the producers of statistics. On the other hand, statistical literature written for users of statistics, literature that addresses the users' needs of interpreting, understanding, and critically examining their data is indeed scarce. *Statistics Explained: Basic Concepts and Methods* mainly directs itself to British readers and provides many insights into the possibilities and restrictions of statistics.

From a pedagogical point of view, the book has an excellent presentation. Each chapter is organized in a systematic fashion. First, a certain problem area is introduced using examples. After that comes a discussion of the "what-how-where" of the data; i.e, a discussion of how the data were collected and the questions that the data were meant to answer. This is followed by a short presentation of the statistical theory used in this particular example. And lastly, some interpretation of the data is reached and this interpretation is evaluated. Each chapter concludes with empirical computer exercises written for Minitab; also there are other exercises and solutions.

Because the theoretical parts are interwoven with a rich amount of text and references, the book also becomes accessible to people who have a restricted knowledge of mathematics. The book thus follows the classic Anglo-American tradition of being both easy to read and of interest for a wide circle of readers. It seems to me that this book should be read by everyone who work in media, such as journalists, and even politicians and researchers in other fields.

My personal reflection is that this book paves the way for a follow-up or companion volume, tentatively named *Making Inferences* with a similar organization and presentation. The question of making inferences is so essential from several aspects that it deserves an entire book and not only an abbreviated chapter as in the present book.

Per Nilsson
Statistics Sweden
Stockholm
Sweden

Matérn, B., Spatial Variation. Lecture Notes in Statistics, 36, Springer-Verlag, Berlin, 1986, ISBN 3-540-96365-0 (Springer-Verlag Berlin), ISBN 0-387-96365-0 (Springer-Verlag New York). 151 pp., DM 33.00.

The first edition of *Spatial Variation* was published in 1960 as *Meddelanden från Statens Skogsforskningsinstitut* 49:5 as Matérn's doctoral thesis. Publication in this form gave it limited circulation, and Springer has now issued a facsimile reproduction of the original together with author and subject indices and a three-page Postscript giving a summary of more recent developments.

It is almost unbelievable how far Matérn was ahead of his time. Some of the theoretical work was started in 1948, and by 1960 Matérn had completed an overview of the theory of two-dimensional random fields and point processes and applied these ideas and those of geometrical probability to sampling problems in forestry. It is this emphasis on sampling, and in particular on the precision of sampling designs, which distinguishes *Spatial Variation* from all subsequent books in spatial statistics. One would expect a 1960 research monograph to be completely outdated by now, but in Matérn's case this is far from so. Matérn quotes only a few additional references on spatial sampling in his postscript, and I am aware of only a handful of others. In part this is testimony to the completeness of his approach.

Chapter 1 is (the original) introduction. The second chapter gives an introduction to stationary stochastic processes on R^n, now more often referred to as random fields. Modern readers may find this difficult. It is stated to be a review, but the material (especially characteristic functions) is no longer emphasized in courses and texts on probability theory. (The first place I encountered a Bessel function was whilst a graduate student reading the first edition!) Section 2.6 sketches the idea for random measures, a subject developed in depth in later years by Olav Kallenberg. In his postscript, Matérn seems to confuse this with the random set theories of Kendall and Matheron, which are quite distinct. The countable additivity which Kallenberg added to Matérn's postulates of finite additivity and second-order stationarity is necessary to avoid a measure-theoretic quagmire. It is also a key part of the reduction of moment measures exploited by Krickeberg and the reviewer.

Chapter 3 discusses some specific mechanisms for constructing stochastic processes. These provide one of the most comprehensive catalogues available of stochastic processes with specified correlation functions, which has proved invaluable in these days of extensive simulation. Later sections introduce some widely used models of point processes and random sets. This chapter contains an amazing richness of ideas, many of which are only now being exploited. The one important idea which is not present is that of Gibbsian point processes (Ripley (1988)). These remove the awkwardness of the "more regular than Poisson" processes of § 3.6.

The fourth chapter begins the more practical half of the study by considering what spatial correlograms occur in practice. Matérn uses a few artificial examples to suggest that an exponential correlation function is appropriate, and then in Chapter 5 considers the efficiencies of stratified and systematic sampling schemes under this correlation function. Since exact calculations were too much for the computers of the 1950s, a number of clever approximations are used. Today the exact calculations can be done without difficulty. The actual values are rather different, but the qualitative conclusions (to use systematic sampling on a rectangular grid) remain unchanged. Chapter 6 is a miscellany of calculations related to practical sampling problems in forestry. I have always found this impenetrable. The topics are only very loosely related and there are few firm conclusions. In part it is a commentary on Matérn's 1947 essay (in Swedish, and even less available than the first edition of *Spatial Variation*).

The field of spatial statistics has been changed radically by the computer revolution, so it is no surprise that it is the more general and theoretical Chapters 2 and 3 which have endured best the passage of 25 years. Indeed, they have never been superceded as a reference for general stationary isotropic random fields. The work on Chapter 5 on sampling plans is still a model of what can be done with simple calculations, and has proved to be an inspiration to the geostatistics school. More extensive experience has suggested that the exponential correlation function is less widely applicable than Matérn implies, and that both border effects and long-range correlations need to be taken more seriously than their easy dismissal here.

The postscript is the weakest part. Perhaps it is unfair to expect Matérn (who was by then retired) to be aware of all the recent developments in spatial statistics, but a good deal of the commentary is ill-informed and lacks the incisiveness of the original material. Much more could be made of the developments in geostatistics (Journel and Huijbregts (1978)) and random processes have been quite widely proposed in the design and analysis of field trials (e.g., Bartlett (1978); Besag and Kempton (1986); Wilkinson, Eckert, Hancock, and Mayo (1983)).

Spatial Variation is certainly of historical interest, and a testament to Matérn's vision. But it is considerably more than that. Although not a suitable introduction to spatial statistics, Chapters 2 to 5 are compulsory reading for anyone with aspirations to specialize in the area. I know very little about forestry applications, but suspect that the wisdom on spatial sampling in this volume is still not widely applied outside Sweden.

References

Bartlett, M. S. (1978): Nearest Neighbour Models in the Analysis of Field Experiments. Journal of the Royal Statistical Society, Series B, 40, pp. 147–174.

Besag, J. and Kempton, R. (1986): Statistical Analysis of Field Experiments Using Neighbouring Plots. Biometrics, 42, pp. 231–251.

Journel, A. G. and Huijbregts, C. J. (1978): Mining Geostatistics. Academic Press, London.

Matérn, B. (1947): Metoder att uppskatta noggrannheten vid linje- och provytetaxering. Meddelanden från statens skogsforskningsinstitut, 36, nr 1.

Ripley, B. D. (1988): Statistical Inference for Spatial Processes. Cambridge University Press, London.

Wilkinson, G. N., Eckert, S. R., Hancock, T. W., and Mayo, O. (1983): Nearest Neighbour (NN) Analysis of Field Experiments. Journal of the Royal Statistical Society, Series B, 45, pp. 151–211.

Brian. D. Ripley
University of Strathclyde
Glasgow
U. K.

Stone, M., Coordinate-Free Multivariable Statistics: An Illustrated Geometric Progression from Halmos to Gauss and Bayes. Oxford University Press, Oxford, 1987. ISBN 0-19-852210-X, xiv + 120 pp., £ 12.00.

This is a nice book and I enjoyed reading it. Multivariate linear statistics is developed using vector spaces and transformations. The book is mathematical. The theory is done in the style of Halmos. A vector space of variables is given. The observations are linear real-valued operators on this space. For example, an observed person NN operates on the variable annual income giving the real number 178 000 SEK. Mean values are also operators on the variable space. The variance is an inner product on the same space. The dual space, where the observations lie, is called the evaluator space.

The book is not directly useful for a practising statistician. He or she will not learn any new methods or learn much about the old ones. However, they will gain some insight into the structure of multivariate linear statistics. They will get a new way of looking at multivariate problems, in particular on the proofs of the basic theorems. A special trait is the use of pictures even for complicated high-dimensional situations. The pictures are sometimes integrated into the proofs of the theorems. The book is probably intended for the graduate level.

The book is in another way quite elementary. It does not require much knowledge of statistics. It does not treat more than the basic theory. For example, principal components, factor analysis, cluster analysis, and discriminant analysis are not mentioned. The book does not even consider estimation when the covariance matrix is not proportional to something known (the Behrens-Fisher problem). The theory can thus not be applied directly to subjects like stratified sampling or nested factors.

I would recommend the book to a mathematician who wants to learn multivariate statistics and who is prepared to read at least one more book in the field. A mathematical statistician, who is used to think of linear statistics in terms of vector spaces, should also benefit from the book and may even enjoy it.

Daniel Thorburn
University of Stockholm
Stockholm
Sweden

Printed July 1989