ACE RESEARCH AND THE CONFIDENTIALITY OF DATA[1]

R. F. Boruch

American Council on Education

The Privacy Issue

Preserving the confidentiality of data has been the subject of much discussion and not a little controversy; numerous articles and books critical or defensive of current policies--have appeared. Westin's (1967) discussion of a wide variety of situations, including wiretaps, commercial records and psychological tests, is a scholarly treatment of privacy jeopardization. More recently, the proposal for a National Data Bank as described by Dunn (1967) and by Sawyer and Schecter, (1968) has led legislators and institutional researchers to give serious attention to the implications of such a project.

Within the social sciences, major concern about the issue has been given formal expression at a number of recent professional conferences. For example, at the 1968 American Personnel and Guidance Association meetings, a symposium, which included representatives of the American Council on Education, the Educational Testing Service, the National Opinion Research Center, and the National Merit Scholarship Corporation, addressed itself to confidentiality of data, specification of respondent rights, and the administrative problems in making data available to researchers. The American Educational Research Association has this year formed a standing committee, which will document and evaluate alternative approaches to existing problems arising from this issue. The American Psychological Association has initiated revision of its current professional code of ethics, based upon questionnaire survey of psychologists. The Russell Sage Foundation (New York) is devoting perhaps the most concentrated attention to the confidentiality question. Results of previous examination of issues under Russell Sage auspices are given by Reubhausen and Brim (1965); Goslin and Bordier are continuing work on educational administrative records.

Two attributes of most of these previous discussions are important in longitudinal research on higher education. First, many of them concern data collection installations which do not function as a source of data for evaluation. Records are used to create judgments about specific individuals. Such records must be highly reliable. Because longitudinal research data frequently has neither the same function, nor the same requirements for reliability, alternative approaches to maintain confidentiality can be developed; the import of this distinction will become evident when the system description is presented.

A second qualification to previous discussions is that most have been confined to justification of social science research in which respondent identification is necessary, and to endorsement of various ethics codes. Rarely do they specify methods for systematically assessing data collection with respect to assurance of confidentiality. Ethical codes are embraced but procedures for implementing these codes are rarely presented.

This paper is intended to clarify some of these issues and to suggest a useful approach to their examination. Using a rather primitive systems analysis, I have attempted to outline the ACE Cooperative Institutional Research Program (CIRP), and the way in which various stages and environments in the program can be assessed relative to the privacy issue. The function of the annual CIRP survey is longitudinal research on biographic, attitudinal and achievement attributes of college students.

ACE Policy and Procedures

Ethical guidelines are useful to the extent that they give to the interested participant or observer an unambiguous acknowledgement of the researcher's concern with safeguarding individual privacy: public misconceptions can be minimized. The delineation of ethics also serves as a useful reference system within which a policy may be implemented administratively.

In view of these considerations, the American Council on Education has formally encouraged the members of its research staff to support the code of ethics adopted by their respective professional organizations. The Council endorses adherence to the codes of ethics of, for example, the American Psychological Association and the National Education Association. When interpreted literally, each professional code includes some brief attention to the privileged nature of the respondent's relationship to the researcher and the obligation of protecting promised confidentiality of that relation. That such codes warrant improvement is apparent from recent professional conferences on the topic.

An explicit statement of policy relevant to administrative records developed and based on these codes of ethics was recommended by the American Council on Education (1967) to colleges and universities. In the statement, formation and implementation of clear policies for insuring confidentiality of students' records were recommended. Guidelines were presented in order to facilitate these recommendations (the use of legal consultation, the elimination of administrative records of political organization membership). Although the recommendations refer to administrative records specifically, they are applied by the ACE Office of Research to survey research records. In addition to these however, ther idealized regulations have been developed which evolve from the functional differences between administrative and survey records. That is, we (a) obtain records anonymously when research objectives do not include merges of data or follow-up studies (b) direct efforts toward safeguarding existing individual records such that identification by anyone (including ACE staff) are minimized or eliminated completely.

To help assure that these policies are implemented within the Office of Research, research activities are monitored by the Council's Board of

Directors, by the Office of Research Advisory Board, and by special advisory committees appointed for specific projects.

Administrative Policy and Procedures

In order to implement a reasonably good protection system, a wide variety of administrative devices have been incorporated into our operations. Regulations can be examined conveniently by using the flow chart given in Figure 1.

At the institutional level of data flow, the respondent group comprises all individuals within the participating institution who provide information about themselves during the questionnaire survey. Personnel at the institution are entirely responsible for administering and collecting questionnaires; the ACE Office of Research furnishes guidelines in order to expedite the process and detailed information is now being made available which describes the nature and functions of the Office of Research, research programs, etc.

The questionnaire contains a statement addressed to the respondent that briefly describes its research function and tells why identifying information is necessary. The respondent is also encouraged to cooperate in the research, under the acknowledgement of ACE responsibility for maintaining the confidentiality of the data. The American Council on Education has no authority to demand that the student respond to the questionnaire, although the institutional authority may indicate that he should complete it.

Under the direction of the institutional representative, the questionnaires are transmitted to a commercial service bureau for optical scanning and magnetic tape record creation. There they are maintained in locked files and destroyed when processing is completed. The product of the optical scanning operation is a statistical file which contains all responses and which uses arbitrary identification numbers for accounting purposes.

In a separate operation, identifying information is recorded on punched cards and then transferred to magnetic tape. This name-and-address file contains only identifying information and an accounting number. If name-and-address files and statistical information were matched, the total file would comprise an "intelligence system," as described by Dunn (1968), for example. However, additional coding at ACE comprises the basis for a double linkage protection system which prevents such matching even by the Office of Research personnel.

The double linkage system (currently being implemented) is illustrated schematically in Figure 2a. Each individual record in a given statistical data file is assigned a unique (arbitrary) accounting number. This series of numerals corresponds to Set 1 in the diagram. Each record in the corresponding name-and-address file is assigned another different accounting number (see Set 2). A code array (CA) of numbers, which match numbers in Set 1 to the corresponding one in Set 2 is created. The code linkage is maintained by a private organization under agreement to (a) allow no direct access to the code system to anyone, including ACE staff, and (b) merge existing accounting numbers with new ones. ACE copies of code linkage are destroyed. In order to implement follow-up studies (Figure 2b) more recent statistical data are assigned new accounting numbers (See Set 3). A new code linkage (CA') is then defined and translated by the service organization to the original system. Merges of statistical data occur without the problems involved in handling statistical and name-and-address files jointly. Accidental or deliberate disclosure of previously collected individual records is impossible simply because the linkage code is not in ACE possession and ACE personnel have no access to them.

At the ACE level of processing, the name-and-address files are maintained at a commercial service organization under the series of administrative constraints given in the Department of Defense Industrial Security Manual (1966) for "CONFIDENTIAL CLASSIFICATION." They are removed briefly from locked storage only for addressing follow-up questionnaires. Accounting controls include receipt and dispatch records, dates and time period of usage, and tape description. These administrative devices are maintained to insure against unauthorized tape copying.

The statistical files are not subject to the same rigid controls prescribed for name-and-address files. These data are so extensive that, even if one had full access to the statistical files and documentation, it would be virtually impossible to match individuals and their responses. Much the same is true for institutional identification.

The statistical data are consolidated and then summarized in various printed forms for the community of users outside the ACE Office of Research. Each institutional representative receives a statistical summary of responses to all questionnaires administered within his institution. This Institutional Report contains no information on individual respondents. To safeguard institutional privacy, ACE refuses to send to an institution another institution's report, although it does advise those researchers and administrators who want to compare their reports to contact each other directly. A second form of report, National Norms for Entering Freshmen (Creager et al., 1968), is a statistical summary of all data for a particular year and is provided to each participating institution and made available to the general public. It identifies neither individuals nor institutions. The only condition under which ACE statistical and identifying data are provided to the institution is met when the questionnaire has been administered under circumstances in which the student has been clearly informed in advance that the data would be returned to his institution for use in local research projects.

Current Problems and Alternative Solutions

At each level of the information system outlined there are potential difficulties in maintaining confidentiality of data. Consider the first level illustrated in Figure 1. Since identifying information appears on each questionnaire, this step represents a reduction in the ACE Office of Research control. That is, students or college personnel may have access to an individual record or a group of records. In cases where most ques-

tions are innocuous or useless for purposes other than research, the threat to the individual is of course, negligible. To the extent that college personnel and students endorse the principle of confidentiality and behave accordingly, the threat is not crucial. Thus far, most institutions acknowledge responsibility to treat document administration and collection confidentially.

More formal provisions for controlling questionnaire administration are possible. Physical security can be enhanced if respondents placed completed questionnaires in locked addressed boxes which would remain unopened until the data processing was initiated. Or the collection, and transmission of documents might be done under the surveillance of local student, faculty, and administrative representatives. More simply, the questionnaire could be constructed so that the identifying information is detachable from that portion of the document on which responses appear, and the identification section and completed questionnaires collected separately. Arbitrary identification numbers imprinted on both documents would permit later collation. The procedure, though not unwieldy, is expensive (approximately one-sixth higher than current processing costs). The types of controls mentioned may generally be too expensive, complicated, or time-consuming to be appropriate in most situations. Perhaps more importantly, very elaborate regulations and procedures could provoke mistrust or suspicion that would interfere with the research or with the operations of the institution. That is, if one implies that suspicion is warranted, then feelings of suspicion may increase or persist unnecessarily. So far, informal surveillance by local administrative personnel seems to be suitable for the general case.

The ACE policy has been to encourage and solicit cooperation by the respondent. The "voluntariness" of the situation is subject to modification by each college. There appears to be some justification for the college's requiring freshmen to complete the questionnaire, in that data are frequently collected in order to provide aid in planning future admissions policies, revising curricula, and other administrative decisions which affect the student. Insofar as the student feels that he cannot conscionably respond to questions relevant to social science or educational research, then conflict will arise. If the justification and restrictions on the research function are made clear to the respondent, conflict can be minimized and the student's cooperation may be forthcoming.

Questionnaire Processing Services

At the second level of the information system, that of the optical scanning of questionnaires and the production of magnetic tape records (name-and-address files and statistical data files), data may be misused, inasmuch as completed questionnaires usually include respondent identification. Irregular data usage at this level could take several forms, the most likely being reproduction of tapes for commercial exploitation.

If safeguarding magnetic tape record confidence is an important objective of the researcher, than endorsement of a relevant code of ethics by professionals in the computing disciplines is a reasonable expectation for reasons described earlier. Although members of the Association for Computing Machinery have discussed guidelines for professional conduct in data processing--some of them relevant to confidentiality--actual results are disappointing. While some members of this community recognize that an explicit code is desirable, other members appear to lack interest, and this indifference is one of the reasons that no code has been adopted. When translatable into procedures for safeguarding privacy of records, such a code can be included into contracts and so strengthen the efficacy of a desirable policy. There are some pitfalls in applying legal principles to technological environments (Banshaf, 1968), but legal applications appear to be a rather basic need if the matter of privacy is considered seriously.

Physical security in the data-processing environments could involve some sort of automated protection for tape or disc files. However, computer manufacturers acknowledge that little effort has been made toward developing devices which have data protection functions. There appear to be three major reasons for this lack of attention (Fanwick, 1967). First, competition among manufacturers is such that the development of such devices is not considered crucial unless there is a substantial demand for them. The demand is low, probably because of the social scientist's naivete regarding use of computing devices on the privacy context. Second, difficulties of undermining current administrative procedures and mechanical devices are sufficient to impede or discourage most attempts to misuse the data. Moreover, the costs required in development of new hardware-software devices may not result in systems which provide more protection than do current systems (Weismann, 1967). Third, it can be argued that in the data-processing environment, records are not available to persons without special skills and that this population of persons competent enough to misuse the data purposely is too small to justify anxiety.

The limited size and nature of the relevant community of potential violators is an important factor, since it means that adherence to existing, commonly used guidelines, can be monitored and controlled well. Further, by including liability and negligence clauses into contracts with individuals or organizations, effective incentives for continued attention to the privacy issue can be provided (Bigelow, 1969).

Office of Research Operations

At the ACE level of the information system, statistical information is analyzed and results disseminated to the public. In addition, follow-up studies are conducted, which require the use of name-and-address files. In addition to those mentioned above, various other safeguards for insuring confidentiality may be considered. The first depends largely on administrative regulation rather than on mechanical or automated procedures. An alternative device relates to computing and data processing methods. Yet another procedure involves capitalizing on the statistical nature of data analysis.

Consider, first, some of the problems that may be inherent in current administration procedures. It may be unwise to invest complete administrative control in a single individual. His personal attributes become too important and his absence, regardless of his character, may cause unnecessary inconvenience in operating the system. A "neutral" organization including members of respondent groups and of the interested professional community, might function as a surveillance or key control unit, in combination with the Director of Research. A second plausible alternative is to extend direct responsibility to other professional staff members of the Office of Research.

The first possibility entails considerable effort, as well as the cooperation of persons outside the ACE organization and so is not possible at present. Both alternatives have the disadvantage of being more bureaucratically complex and time-consuming than the current method.

The second device might involve an in-house computer system which permits merges of data while completely denying any individual direct access to internal name-and-address files and assigned accounting numbers. A model operation would include controls to eliminate direct handling of existing name-and-address files. Matching may be conducted independent of existing or newly obtained statistical files. The merge operations which combine new and existing statistical data are based on the accounting system associated with the existing name-and-address files. If such a system indeed prevents complete access during operations, its importance cannot be underestimated. By making direct linkage of name-and-address information and statistical data impossible, no attempt (legal or otherwise) could possibly threaten the confidentiality of the records and thus the privacy of a single person would no longer be an issue. However, a hardware-software system of this type is not currently available, although some systems can be developed to partially meet requirements. Relative to current administrative and mechanical devices, systems which completely eliminate accessibility would be considerably more expensive.

Consider now the third alternative, that which involves the form of the statistical data on which analyses are based. Typically, the researcher attempts to maintain an isomorphic relation between a person's responses on a questionnaire and records of these responses transformed to magnetic tape form. Now, the possibility of data use or misuse is, of course, weakened when data are not reliable for any <u>specific</u> individual record. Frequently, the researcher can afford to undermine deliberately the integrity of a single record but preserve the integrity of the whole, at least with respect to statistical parameters. He can do so by innoculating statistical data files with randomized error whose properties are known. A large body of literature deals with the problem of adjusting statistical estimates of population parameters, when the observations are subject to <u>known</u> measurement error. The innoculation accomplishes a number of important objectives. First, in the context of public interest in survey research, confusion between administrative records, eavesdropping devices, intelligence systems etc., may be minimized. The controlled unreliability of any individual record is a notion that can be communicated to the public. Second, the likelihood that records will be used in formation of judgments about specific individuals is reduced substantially. One cannot obtain unambiguous information about a specific person, even if identification is, infact, accomplished.

The procedure is inappropriate, of course, for detailed administrative records, but in the survey environment we are not so restricted. For many survey researchers within specific substantive areas (education, psychology, sociology) basic techniques (e.g., regression analysis, discriminant functions) for examining the structure of data can be augmented by including information available on measurement error parameters. Standard computer programs can be altered rather easily to adjust parameter estimates. Although nonparametric approaches which include misclassification probabilities not too well articulated, the current work on the topic ought to allow survey researchers some further latitude (e.g., Assakul and Procter, 1965).

Observations on the Legal Environment

Conceivably, public or private investigatory agencies may have an interest in an individual's responses to an ACE Office of Research questionnaire. Frequently, this information is already public. The researcher in possession of identical information is not confronted with a problem; the investigatory agency can obtain the data more efficently and easily from the public sources. The ACE Office of Research usually asks rather <u>general</u> information of the student, a strategy which not only protects against outside interference but also minimizes the intrusiveness of the questions. When solicited information is nonspecific, the risk of possible harassment to respondent and researcher is minimized but not eliminated.

Although the possibility that specific records survey data will be subpoenaed is a frequently mentioned bugaboo, the likelihood of this event actually happening is rarely assessed. Individuals who perceive such a threat frequently do so on emotional grounds rather than through systematic examination of the particular survey circumstance. It should be pointed out that (so far) in no instance has a subpoena of behavioral research survey data been effected. This lack of legal precedent can be interpreted in several ways. First, individual records directly relevant to investigatory objectives can usually be obtained easily through other agencies. Thus, an investigatory group has no need to solicit information which is of dubious relevance to its interests and which involves direct interference in social science research. The original survey questionnaires contain no signatures, and records of such documents are subject to well known psychometric limitations (i.e., unreliability of responses, and inaccuracy of data processing and of information transmission). In short, the survey research record is unlikely to have any value for use in litigation against a specific respondent.

The sociolegal aspects of the confidentiality issue are interesting and frequently paradoxical. Indeed, only recently has the research participant's right to privacy of personality been acknowledged by judicial and legislative action (see Reubhausen

and Brim, 1965). For the researcher, the so-called privileged communication laws for psychologist-client relations appear to be relevant, but these regulations are statutory (extant in only 18states) and have been applied in rather limited situations.

Even though legislative or judicial inquiry into individual records is unlikely and even though strategies are employed to minimize private or public interest in such inquiry, the need for a legal definition of rights and obligations remains. Reubhausen and Brim, in a detailed examination of these questions, have offered suggestions for their resolution: One is that privileged status be extended to information acquired by the social scientist, another, that civil or criminal remedies for breach of the right of privacy be provided. Unfortunately, none of these suggestions are likely to receive immediate attention, evaluation, and action by the courts or legislative agencies. For the time being, heavy reliance must be placed on ethical codes, and on administrative procedures for their implementation within the particular survey condition. To these ends, the social scientist must devote serious effort to evaluating and implementing alternative strategies at each level of the information system.

References

Assakul, K. and Proctor, C.H. Testing Hypothesis With Categorical Data Subject to Misclassification. Institute of Statistics, Mimeograph Series No. 448, North Carolina State College, Raleigh, North Carolina: September, 1965.

Banzhaf, J.F. "When Your Computer Needs a Lawyer." Communications of the ACM. Vol. 2, No. 8, August 1968.

Bigelow, Robert P. "Some Legal Aspects of Commercial Remote Access Computer Services." Datamation. Vol. 15, No. 8, pp. 48-52. Chicago, Illinois: August, 1969.

Creager, J.A., Astin, A.W., Boruch, R.F., Bayer, A.E. "The National Norms for Entering College Freshmen-Fall 1968." ACE Research Reports. Vol. 3, No. 1, Washington, D.C.: American Council on Education, 1969.

Dunn, E.S. "The Idea of National Data Center and the Issue of Personal Privacy." The American Statistican. Vol. 21, No. 1, February, 1967, pp. 21-27.

Fanwick, Charles. "Computer Safeguards: How Safe Are They?" SDC Magazine. Vol. 10, 1967, pp. 26-28.

Ruebhausen, O.M., and Brim, O.G. "Privacy and Behavioral Research." Columbia Law Review. Vol. 65, 1965, pp. 1184-1211.

Sawyer, Jack, and Schechter, Howard. "Computers, Privacy, and the National Data Center: The Responsibility of Social Scientist." American Psycholgist. Vol. 23, No.11, November, 1968.

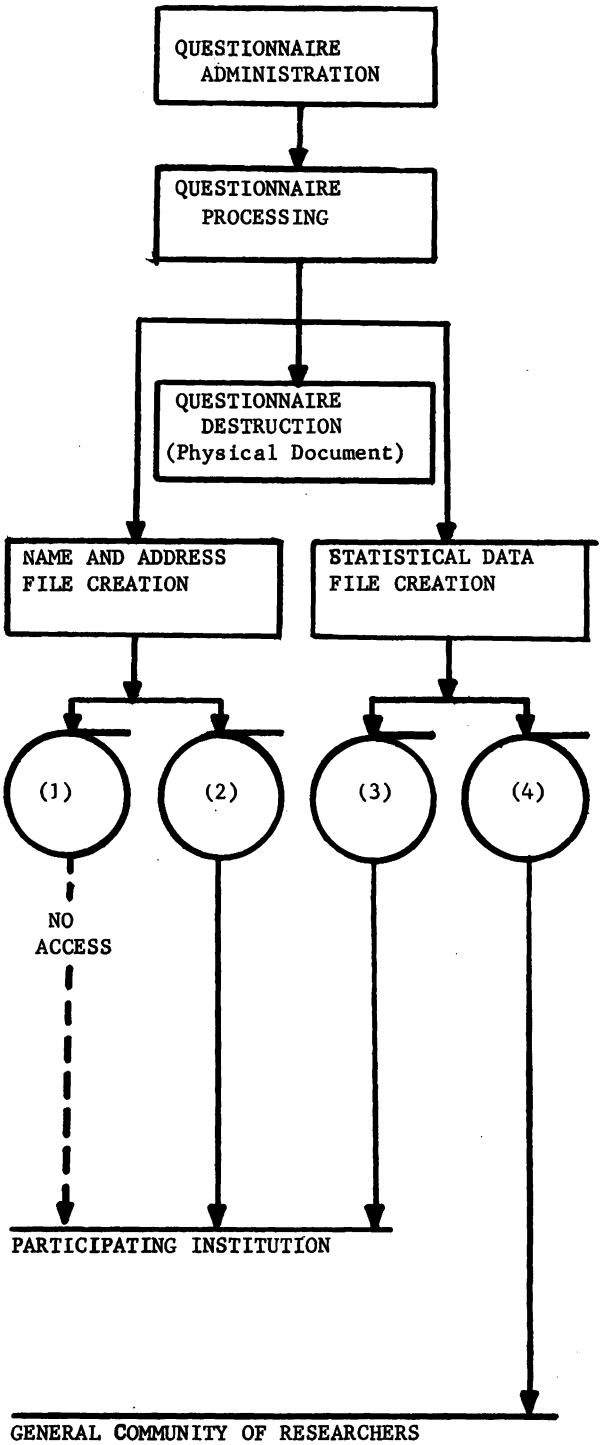
U.S. Department of Defense. Industrial Security Manual for Safeguarding Classified Information. DOD 5220, 22-M. Washington, D.C.: Department of Defense, July 1, 1966.

Weismann, Clark. "Programming Protection: What Do You Want to Pay?" SDS Magazine. Vol. 10, No. 7, July, 1967, pp. 30-31.

Westin, A.F. Privacy and Freedom. New York: Atheneum, 1967.

Footnote

1. Partial support for this research was provided by NIMH Grant NM17084-01. Opinions expressed are not necessarily those of the sponsoring agency.



(1) DATA MERGE OPERATIONS
(2) STUDENT NAME AND ADDRESS FILE
(3) STATISTICAL (within college) REPORT
(4) NATIONAL NORMATIVE REPORTS

FIGURE 1. INFORMATION FLOW

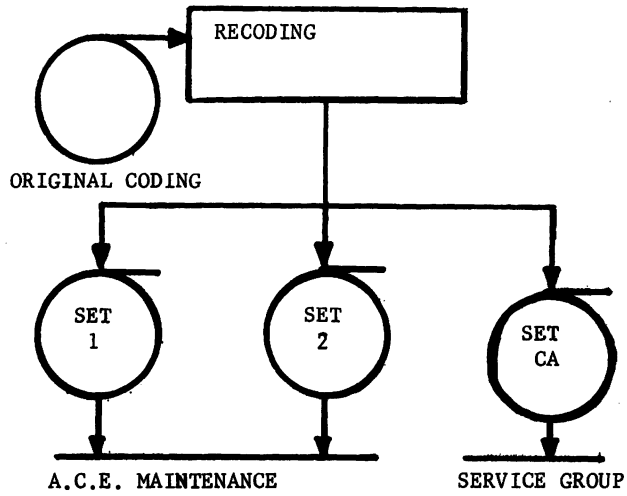


FIGURE 2a. LINK FILE CREATION

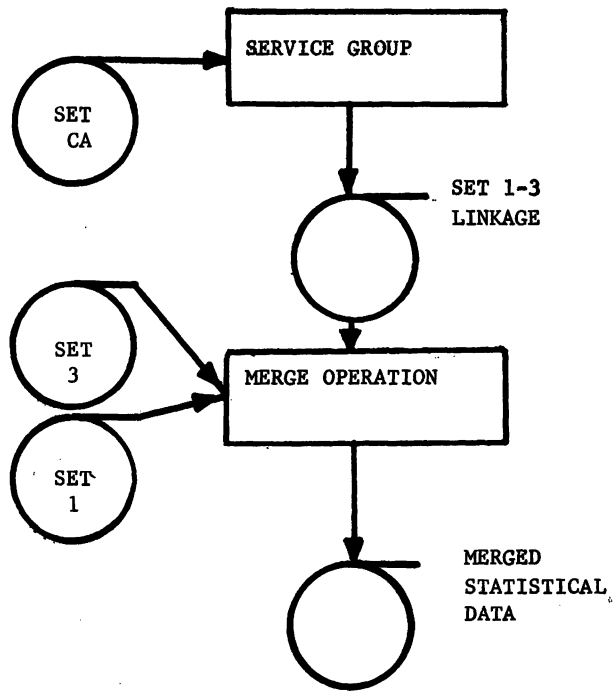


FIGURE 2b. MERGE PROCESS