Robbin, A., & Koball, H. (2001). Seeking explanation in theory: Reflections on the social practices of organizations that distribute public use microdata files for research purposes. *Journal of the American Society for Information Science*, *52*(13), 1169-1189.

Seeking Explanation in Theory: Reflections on the Social Practices of Organizations that Distribute Public Use Microdata Files for Research Purposes

Alice Robbin
School of Library and Information Science, Indiana University, 005 Main Library, Bloomington, IN 47405. E-mail: arobbin@indiana.edu

Heather Koball
Urban Institute, Washington, DC. E-mail: hkoball@ui.urban.org

Public concern about personal privacy has recently focused on issues of Internet data security and personal information as big business. The scientific discourse about information privacy focuses on the crosspressures of maintaining confidentiality and ensuring access in the context of the production of statistical data for public policy and social research and the associated technical solutions for releasing statistical data. This article reports some of the key findings from a small-scale survey of organizational practices to limit disclosure of confidential information prior to publishing public use microdata files, and illustrates how the rules for preserving confidentiality were applied in practice. Explanation for the apparent deficits and wide variations in the extent of knowledge about statistical disclosure limitation (SDL) methods is located in theories of organizational life and communities of practice. The article concludes with suggestions for improving communication between communities of practice to enhance the knowledge base of those responsible for producing public use microdata files.

> ***Exemplary*** **survey research practice requires that one literally do "whatever is possible" to protect the privacy of research participants and to keep collected information they provide confidential or anonymous. (American Association of Public Opinion Research, 1999, p. 5)**

Introduction

More than 30 years of public opinion polls record significant concerns about the quantity and use of personal information collected by government and the private sector and computer technology whose uses are perceived to diminish personal privacy.[1] Public anxiety fueled the implementation of federal and state statutes, regulations, and administrative policies and guidelines to safeguard privacy and create enforceable expectations of confidentiality in the United States, as well as strong data protection laws in Canada and Western Europe.[2]

Public attention has recently focused on three intersecting issues: the practices of the U.S. Bureau of the Census and the 2000 decennial census, Internet data security, and personal information as big business. The 2000 decennial census became a highly politicized issue because it requested personal information that was deemed intrusive.[3] Lapses in data security have heightened fears that personal information cannot be protected, and, at the same time, there is growing public recognition that there are few legal protections for personal information owned by the private

Received November 27, 2000; Revised February 19, 2001; Accepted April 30, 2001

[1] See Louis Harris Associates and other public opinion data on this subject, available from the Institute for Research in Social Science, University of North Carolina at Chapel Hill at: http://www.irrs.unc.edu/. For a summary of 30 years of public opinion polls, see Robbin (2001a).

[2] The U.S. Privacy Protection Study Commission's (1977) publication, *Personal Privacy in an Information Society* remains the seminal document for evaluating government's role in personal and information privacy. Alan Westin should be acknowledged as a key agent in alerting the public to the risks of computer technology and loss of personal privacy (Westin, 1971, 1976a, 1976b; Westin & Baker, 1972). Western European countries have strong data protection, which the United States does not have (see Bennett, 1997, 1998; Bennett & Raab, 1997; Cate, 1997; European Commission, 1997; Perrine, 2000).

[3] There was extensive press coverage of the decennial census and its intrusiveness (e.g., Norman, 2000; Rosenbaum, 2000), which was also examined by various researchers, including Martin (2000), Nie & Junn (2000), Nie, Junn, and Slotwiner (2000), and Robbin (2001a, 2001b).

sector.[4] Indeed, the American Society for Information Science and Technology's (1997, p. 2) *Bulletin* has, for example, reflected on the "difficulties in maintaining both the integrity of our data and personal and corporate privacy [that] are growing by orders of magnitude."

Politics, law, and policy on privacy and confidentiality, Internet security breaches, and the misuse of personal information by e-commerce firms have also had far-reaching consequences for government's statistical activities and the social science enterprise.[5] Statutory protections with civil and criminal penalties make government agencies and researchers accountable for protecting information about human subjects and corporate entities.[6] At the same time, however, law and policy recognize that individuals do not have "complete authority to control whether and on what terms they participate in [civil society, corporate life], and socially important research" (Mackie & Bradburn, 2000, p. 20). Statutes and executive agency regulations also permit the release of data under restricted use conditions that protect the confidentiality of records, when justified as necessary for the conduct of scientific discovery and for informed public policy.[7]

The scientific discourse about the issues of privacy, confidentiality, and the release of data addresses the cross pressures of maintaining confidentiality, and ensuring access in the context of the production of statistical data for public policy and social research (Cox, 1991; Duncan, Jabine, & de Wolf, 1993; Duncan & Pearson, 1991a, 1991b; Kruskal, 1981; Norwood, 1991; Fienberg, Martin, & Straf, 1985). Maximizing the two privileged but competing social values of data protection and data access has led to a large corpus of published research on statistical disclosure limitation and confidentiality-preserving statistical methods by statisticians in Western Europe and North America. This research, some of which is discussed in this article, has made it possible to release public use files that serve as the foundation for social science research and public policy.

The decision-making process for producing public use files that preserve confidentiality remains, however, more or less invisible to the general public. We lack empirical evidence about the cross pressures experienced by technical staff and the ways in which the public debate on privacy and confidentiality influences the technical decisions about the production and release of public use files. What the rules regarding disclosure risk mean and how they are used by people in the context of bureaucratic work have generally been ignored as an empirical issue. This article seeks to fill a gap in our knowledge base about organizational decisions to produce public use files that contain person information and also contributes to a small corpus of research that investigates how "back office," mostly invisible information technology workers do their work (Clement, 1994; Star, 1991; Suchman, 1996).

We report findings from a small-scale survey of the methods used by government agencies and private sector survey research organizations to limit disclosure of confidential information prior to publishing public use microdata files and the cross pressures that members of these organizations experienced in releasing or denying access to statistical data.[8] A key finding of this organizational practices survey was that people responsible for preparing data files that contained potential risks for disclosing the identity of subjects appeared to have taken few precautions for limiting disclosure. This finding was unexpected, because extensive research has been carried out to devise techniques that minimize risks, and a large published literature on the subject exists.

The finding led the authors to a central research question: "What prevented these organizations and their information managers from having received 'state of the art' information about statistical disclosure risk and methods to limit that risk." Put another way, why does knowledge "out there" not get used and incorporated into organizational practice? The explanation employs theories of organizational behavior and communities of practice to explore a variety of organizational contingencies, especially those internal to organizations, such as local work practices, cultures, complexity, and other barriers to useful and important statistical data.

[4] U.S. Department of Commerce (1999); U.S. Federal Trade Commission (1998, 2000); Clausing (2000); Americans facing loss of personal privacy (2000); Frishberg (2000); McCullagh and Morehead (2000); Caruso (1999); Okie (2000); Mosaquera (2000).

[5] They also have had significant effects on private and corporate decision making. Martin's (1998) study of employment trends in the information services and technology sectors shows indirectly how corporate decisions about information technology infrastructure building, R&D investments, expansion of training programs, wage and salary structures, and global competitiveness depend on the analysis of statistical data that are collected at the level of the individually identifiable firm and individual. Martin's analysis rests on data that are released as "public use" (that is, made available to the public) "microdata" or summary tabulation (aggregated microlevel data) files.

[6] For example, the Privacy Act of 1974, the Computer Security Act of 1987, and the National Education Statistics Act of 1994 direct the National Center for Education Statistics and their contractors to protect confidential data (McMillen, 1999).

[7] One example of administrative rules that permit the release of individually identifiable information pertains to the Health Care Financing Administration. "[Medicare entitlement, utilization, and payment data] may be released . . . (4) To an individual or organization for a research, evaluation, or epidemiological project related to the prevention of disease or disability . . . if HCFA: . . . (b) Determines that the purpose for which the disclosure is to be made: (1) Cannot be reasonably accomplished unless the record is provided in individually identifiable form; (2) Is of sufficient importance to warrant the effect and/or risk on the privacy of the individual that additional exposure of the record might bring . . ."(U.S. Department of Health and Human Services, 1992, p. 53763).

[8] This survey was originally carried out to provide background information for an October 1999 workshop sponsored by the Committee on National Statistics (CNStat) of the National Research Council (Mackie & Bradburn, 2000; Robbin & Koball, 2000; Robbin, Koball, & Jabine, 1999). The summary of recommendations that resulted from our survey is published in Mackie and Bradburn (2000). We would like to acknowledge Thomas Jabine, who has played a central role in federal statistical activities regarding data confidentiality and data sharing and served as consultant for this survey.

Table 1. Statistical disclosure limitation techniques and definitions.

Statistical disclosure limitation techniques	Definitions
Collecting or Releasing a Sample of Data	Releasing subsamples of larger data sets. For example, the Census Bureau releases a sample of census data as public-use microdata.
Including Simulated Data	
Multiple Imputation[1]	Using multiple imputation techniques to create a simulated data set with the same statistical properties as an original data set.
Blurring Data, Grouping, or Adding Random Error	
Top-Coding	
Bottom-Coding	Grouping extreme high values into one category.
Intervals	Grouping extreme low values into one category.
Noise Introduction	Grouping continuous, unique values into broader intervals
Blurring	Introducing random error into data
Microaggregation	Replacing a reported value with an average value for a subgroup. Grouping records based on a proximity measure of all variables of interest, and calculating aggregates for those variables.
Excluding Certain Attributes	
Deleting Identifiers	Deleting unique identifiers, such as day of birth.
Dropping Sensitive Variables	Dropping particularly sensitive variables, such as HIV-status from public-use microdata sets.
Blank and Impute	Selecting records at random, blanking out selected variables and imputing for them.
Swapping Data	
Record Swapping	Selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all other variables.
Rank Swapping	A way of using continuous variables to define pairs of records for swapping; instead of variables matching exactly, they can be close.

Sources: Duncan et al., 1993; U.S. Federal Committee on Statistical Methodology, 1994, July 1999.

[1]Imputation refers to procedures for assigning values to unreported items, as a way to minimize nonresponse, inconsistent, or unacceptable codes. The same procedures may be used to protect confidentiality. For more information on applications, see U.S. Bureau of the Census publications, such as the Technical Documentation released from the 1990 decennial census (Appendix C. Accuracy of the Data) available at: http://www.census.gov

Part 1 introduces the concepts of public use microdata files and statistical disclosure limitation (SDL) methods, and provides examples of how disclosure risk is evaluated. The section serves both to show the ways in which disclosure risk is evaluated by statisticians and to provide the supporting evidence for our conclusion that most of the organizations we surveyed did not employ SDL procedures prior to releasing longitudinal public data files. The section is also useful for database designers and developers of data mining tools, who work with large amounts of personal information. Part 2 explains how the Survey of SDL Organizational Practices was conducted. The strategy for drawing the sample of microdata files is discussed, but the names of data files are not revealed to protect the confidentiality of the respondents and their organizations. Part 3 reports how the rules about preserving confidentiality were applied in practice by our respondents, including how their decisions contributed to disclosure risk, how they interpreted the rules about maintaining confidentiality of personally identifiable information, and how the environment they operated in contributed to a general lack of knowledge about SDL techniques.

Part 4 discusses the findings in the context of seeking an explanation for the apparent deficits and wide variations in the extent of knowledge about SDL methods. We apply theories of work life in large-scale complex organizations and the role that knowledge plays in creating and sustaining communities of practice. We conclude by restating how theory was critical to understanding organizational communication failures and also offers guidance for improving communication between communities of practice to enhance the knowledge base of people responsible for the production of public use microdata files. Tables 1 and 2 identify the major statistical disclosure limitation techniques and definitions, and administrative procedures for restricting access to microdata files and definitions, respectively. Appendix 1 contains a facsimile of the SDL Organizational Practices Survey questionnaire. Appendix 2 describes the characteristics of and conditions for access to the microdata files that were part of this survey.

Reducing Disclosure Risk in Public Use Longitudinal Microdata Files

Public use files distributed by government and the social research community are considered a major scientific innovation that has proved to be an effective solution for balancing the "often conflicting goals of preserving confidentiality and exploiting the research potential of microdata" (Mackie & Bradburn, 2000, p. 1). The demographer Stephen Ruggles has written an assessment of the best known of the public use microdata files, the 1940 through 1990 decennial census Public Use Microdata Survey (PUMS) files, that can be extended to other public use microdata files: PUMS files "revolutionized the analysis of the American population;" and led to "an explosion of census-based

Table 2. Administrative procedures for restricting access to microdata files and definitions.

Administrative Procedures	Definitions
Special Employee Arrangements	Giving a person who does not work for an agency or organization special status to access confidential data, such as having special sworn-in status.
Restricted Locations of Access	Restricting data access and analysis to an on-site location.
Modes of Data Transmittal	Releasing encrypted data with software that can produce descriptive statistics from the encrypted data.
Off-Site Physical Security Measures	Requiring off-site data users to safeguard access to restricted data.
Oaths and Written Agreements	Signing an agreement between the data collection organization or agency and the data user. This agreement often specifies the intended use of the data and the individuals who have access to the data.
Penalties	Imposing penalties for the release of confidential data. Penalties can range from denying further access to data to imprisonment for the release of Title 13 census data.
Disclosure Review Board	Formally constituted agency unit tasked with evaluating the risk of disclosure, comprised of representatives of the department's statistical staff, subject specialists, and, for National Center for Education Statistics, a representative of the Bureau of the Census. Producers of public use files complete a form that provides general rules about geographic information and the contents, in order to determine whether there is an unusual risk of individual disclosure.

Sources: Jabine, T. B. (1993); McMillen (1999, p. 3). U.S. Federal Committee on Statistical Methodology (1999).

research" and to the PUMS as "the mainstay of American social science" (Task Force, 2000, p. 1).

Many complex policy issues can only be understood with longitudinal or panel survey data (Boruch & Pearson, 1985). Mackie and Bradburn (2000) discuss how these types of data "facilitate a broad spectrum of research [that] shed[s] light on . . . questions that could not otherwise be reliably conducted" (p. 6) when they are linked to "health, economic, contextual [e.g., unemployment rates], geographic, and employer information" (p. 2).[9] They also explain that linking survey data and administrative records, such as social security earnings, unemployment compensation, medicare benefits, hospital records, or student transcripts, can be an efficient solution to the high costs of data production because linkage "reduce[s] the need to duplicate survey activities, reduce data collection and processing, and improve data quality" (p. 6).

Individual-level sample data from the decennial census, as well as other national probability samples, which represent a very small fraction of the population can be released after sensitive data are deleted, and minimal detail on geography is provided to protect the confidentiality of respondents. Statistical methods designed to limit disclosure risk and administrative rules that provide access under controlled conditions offer a theoretical framework and practical solutions, respectively, for balancing these conflicting values. Tables 1 and 2 summarize some of the solutions applied by data producers to release or permit access to microdata files, and general approaches to reducing disclosure risk are discussed in the next section.

The stewardship of statistical data for policy and research remains a problematic issue, however. At the same time that public use microdata files, including longitudinal and panel studies, have become an essential tool for conducting scholarly research and public policy, their release always entails some risk of disclosure and, therefore, potential harm to respondents.[10] There never exists a condition of zero risk, which statisticians recognize as "an impossibly high standard"; and, moreover, "collection of any data entails some sort of risk, no matter how small" (American Statistical Association, 1999a, p. 1; U.S. Federal Committee on Statistical Methodology, 1994).

[9] For examples of this type of research, see Duncan and Pearson (1991). There is a long history of discussion about the contribution that social science makes to public policy, some of which is available in Weiss (1977a, 1977b); National Research Council (1979); National Science Board (1969); McAdams, Smelser, & Treimar (1982); U.S. House Committee on Government Operations, Legislation and National Security (1982); and Martin and Straf (1985).

[10] The ways that disclosure of a person's identity may occur are summarized by the U.S. Federal Statistical Committee on Methodology (1994, p. 1): "Disclosure occurs when information that is meant to be treated as confidential is revealed. Sometimes disclosure can occur based on the released data alone; sometimes disclosure results from combination of the released data with publicly available information; and sometimes disclosure is possible only through combination of the released data with detailed external data sources that may or may not be available to the general public."

Statisticians have demonstrated that access to the records of the original population (or universe) creates the conditions for verifying identification of members of the sample, although identification requires a high level of computational and statistical literacy to verify or confirm the identity of a particular individual (Bethlehem, Keller, & Pannekok, 1989; Kim & Winkler, 1995; Winkler, 1997, 1998). Releasing a sample as a public use file and making unavailable the original population/universe data greatly reduces the risk of disclosure, although disclosure is never completely eliminated because full information on all possible risks is not available (see Simon, 1979/1955 on "bounded rationality"). We do know, however, that the possibility for *inadvertent* disclosure is magnified when these files contain detailed characteristics of respondents. Identity also becomes more vulnerable to disclosure when data for respondents who participate in long-term research investigations are linked to administrative records of government agencies.

Yet it is this detailed information, longevity of data collection, and the linkage to other types of information that enhance the research potential of the data. Conversely, the less data accessible for analysis, the lower utility for research and public policy. Thus, the core issue for data distributors of public use files is: how much risk will be tolerated, so that public use files protect confidentiality but are not too stringent to constrain the utility of the data.

General Approaches to Reduce Disclosure Risk

Two general approaches, *restricted data* and *restricted access*—statistical techniques and administrative procedures, respectively—have been utilized to protect the privacy and confidentiality of respondent information and provide access to statistical data (Duncan, in press; Duncan et al., 1993; Feinberg & Willenborg, 1998; Sullivan, 1992; U.S. Federal Committee on Statistical Methodology, 1978, 1994, 1999).[11]

Restricted data applies constraints on access to the *content* of data sets or files to be released in microdata files. Content is restricted by such methods as removing explicit identifiers; reducing information by creating intervals, grouping or bounding permitted values (top or bottom coding); recoding variables; deleting sensitive variables; suppressing values in a cell; injecting random error; or releasing only a sample of the data (e.g., 18% sample of the 1990 Census). These techniques are called "data conditioned methods" (Duncan, 2001, p. 9). Another type of disclosure limitation method substitutes one sample of data for another, known as "synthetic" data (see Kennickell, 1999a, 1999b); although the risk of identity disclosure is eliminated with the latter approach, the risk of attribute disclosure may be possible (Duncan, 2001, p. 10). Table 1 summarizes the various methods and their definitions.

Restricted access imposes "conditions on who may have *access* to agency [or organization] data, for what purpose, at what locations, and so forth" (Duncan et al., 1993, p. 142). Government agencies recognize that compliance with confidentiality requirements might severely curtail the "amount of detail included in statistical summaries or microdata sets intended for unrestricted access, [and] such limitations inevitably limit the scope and depth of analytical uses of the data" (Jabine, 1993, p. 538). Specified written and enforceable contractual agreements define the responsibilities of the researcher. Table 2 summarizes these procedures and their definitions. For more discussion on restricted access procedures that government agencies have applied, see the summaries in Jabine (1993), Mackie and Bradburn (2000), and the American Statistical Association (1999b).

Example of Assessment of Statistical Disclosure Risk

The evaluation of disclosure risk in the *National Education Longitudinal Study of 1988* (NELS) illustrates the steps that analysts took to reduce disclosure risk prior to developing public use files for cross-sectional and longitudinal surveys. The NELS provides both public use and restricted use files that are longitudinal surveys of high school students. Some of the data in the public use files have been altered or suppressed, whereas the "restricted use" files "preserve the original data free of all confidentiality edits" (Ingels, Dowd, Baldride, Stipe, Bartot, & Frankel, 1994a, p. 1). The latter are available through the National Center for Education Statistics data licensing system (see McMillen, 1999).

This example illustrates the statistical methods that are available, but it is also important to recognize that application of appropriate statistical disclosure limitation (SDL) methods is a function of the type of data, their detail, and anticipated uses because a general set of rules governing releases is not possible. Jabine (1993, p. 542) comments, "Virtually every proposed release of a new data set differs in some relevant way from other releases, even within the same agency and program (see also Mackie & Bradburn, 2000, p. 41).

The NELS statistical program reflects the mandate of the Department of Education National Center for Education Statistics (NCES) to "collect and disseminate statistical and other data related to education in the United States" and "to conduct and publish reports on specific analyses of the meaning and significance of these statistics." The NELS study is part of a program that now spans three decades. The program began in the early 1970s with a cohort of students in the *Longitudinal Study of the High School Class of 1972* (NLS-72), which was then followed by a second cohort designed to study developments during the 1980s, *High School and Beyond* (HS&B). NELS examines developments during the 1990s and beyond that, and is designed to

[11] This discussion on statistical disclosure risk focuses only individual units of analysis that are individual persons or households. There are important and very difficult issues concerning confidentiality for establishment (firm)/institutional units because there are fewer establishments, publicly available data, and greater motivation to identify firms (see Duncan & Pearson, 1991b, p. 238; Duncan et al., 1993; Norwood, 1991).

"develop and examine federal education policy," with data provided on the "educational, vocational, and personal development of students at various grade levels, and the personal, familial, social, institutional, cultural factors that may affect development" (Ingels et al., 1994a, pp. 7, 1–2). Public use files from all three cohorts have been very widely disseminated, and the corpus of research is substantial.

NELS is a nationally representative sample of eighth-grade schools and students achieved by a two-stage stratified probability sample design (Ingels et al., 1994a, p. 7).[12] The technical report explains that NELS was designed "to address multiple research and policy objectives" (p. 6). The "major features" of the study are "supplementary components to support analyses of geographically or demographically distinct subgroups" and "linkages to previous longitudinal studies and other current studies" (p. 6).

Three types of analyses can be carried out: "cross-wave, cross-sectional at a single point in time, and cross-cohort by comparing the NELS cohort to cohorts in the NLS-72 and HS&B" (Ingels et al., 1994a, p. 6). The longitudinal ("cross-wave") feature provides researchers with the capability of studying the students' educational and professional attainment. The cross-cohort capability means that "comparisons can be drawn with previous NCES longitudinal studies" (p. 7), and trend analysis can be carried out. Some of the NELS content is also contained in the NLS-72 and the HS&B studies. A core set of data is continued, but NELS is also designed to incorporate new research and policy questions (see Owings, 2000, as well as other NCES publications on the objectives of the NELS study).

NELS includes assessment and transcript records on the students and also follows students who drop out of school. The longitudinal data are "augmented through parent, teacher, school administrator (questionnaires were administered), and school record accounts of students' progress and development," that are designed to understand their "problems and concerns, relationships with parents, peers, and teachers, and the characteristics of their schools" (Ingels et al., 1994a, p. 6).

The technical reports summarize the extensive disclosure risk analysis of the NELS data. The researchers identified "items of information, used alone, in conjunction with other variables, or in conjunction with public external sources such as school universe files" to ensure that no institution's or individual's identity could be disclosed (Ingels et al., 1994a, p. 13). Variables were suppressed or altered if they posed significant disclosure risks (p. 14). A two-step evaluation took place to examine disclosure risk, first, on each of the cross-sectional files and, then, on the longitudinal data (base data and followups). For example, for "an extremely small number of schools" (p. 113), eight variables were suppressed, including race, ethnicity, region, any value over 10 for family size (recoded to 10). Thirteen parent component variables were also altered, including identifiers, residential location, race, ethnicity, region, and family size (fn, p. 113).

Cross-sectional Disclosure Risk Analysis

The researchers "pre-identified those variables deemed high risk because they 'constituted virtually unique data signatures,' such as continuous variables, extreme outliers, and 'finer-grain versions of school-level'" information (Ingels et al., 1994b, p. 113). High risk continuous variables were recoded as categorical variables. Detailed categorical variables were recoded into larger categories. An institution or student might be coded as "missing, coded to an adjacent category, or included in a code which collapsed two or more response categories" in the public use files, if either the institution or student could be "characterized in terms of a single variable in the original data" (Ingels et al., 1994a, p. 14).

The researchers analyzed potential disclosure risk for school level information because school-universe files are publicly available (Ingels et al., 1994b, pp. 114–115). The "Common Core of Data" (CCD) is the Department of Education's primary database on public elementary and secondary education in the United States, containing comprehensive information for approximately 90,000 schools and about 16,000 school districts (approximately 16,000).[13] Quality Education Data (QED) is a for-profit research and database company, whose "National Education Database" covers U.S. and Canadian educational institutions. QED also maintains a "National Registry of Teacher Names"

[12] NELS has five components: a sample of students who were eighth graders in 1988 and are followed up at 2-year intervals. The *N* for the original sample (base year) of participants is 24,599 (Ingels et al., 1994a, p. 7; 1994b, p. 116). A total number of 1,252 schools contributed usable data in the base year 1988. The original sample is augmented in subsequent data collections to maintain the national representativeness of the grade-specific cohort (e.g., the original sample of eighth graders who are now sophomores is representative of a national cohort of sophomores). The study now incorporates the base data collected in 1988 and four follow-ups (1990, 1992, 1994, 2000). For more information on the NELS, see http://nces.ed.gov/surveys/nels88/.

[13] The quantity of publicly available information about elementary and secondary schools is so extensive that it greatly increases disclosure risk and, thus, accounts for the very careful disclosure risk assessment that the Ingels team carried out. The U.S. National Center for Education Statistics annually collects detailed information on all public elementary and secondary schools in the United States, which originate from five surveys sent to the 50 state education departments. The information is then published in the NCES's "Common Core of Data" (CCD) (for more detail, see http://nces.ed.gov/ccd/aboutccd.html). For example, the school district level information includes phone number; location and type of agency; current number of students; and number of high school graduates and completers in the previous year. State-level aggregated information includes the number of students by grade level; full-time equivalent staff by major employment category; and high school graduates and completers in the previous year. State-level fiscal information is also reported, including average daily attendance; school district revenues by source (local, state, federal); and expenditures by function (instruction, support services, and noninstruction) and subfunction (school administration, etc). The school district level fiscal data describe revenues by source and expenditures by function and subfunction, and enrollment.

database of teacher, curriculum and school data "gathered directly from teachers and administrators."[14]

The analysts identified 10 variables that were in both the NELS and QED database and seven variables in both the NELS and CCD. They constructed a "code distance metric" to calculate the distance between schools on selected variables.[15] If their analysis of school matches showed that institutional identity could be deductively disclosed, then they made more changes to school-, student- or teacher-level variables (p. 113). They also had to assure that the "abridgements, recategorization, and maskings made for confidentiality purposes on school data" were applied to the student records (p. 113).

The NELS:88-QED assessment yielded 98 schools that were at risk of disclosure, or 7.8 percent of the sample schools (98/1252). The researchers then applied methods to further reduce the risk of disclosure "to an acceptable level" (Ingels et al., 1994b, p. 114). They removed the "percent Black" and "percent Hispanic" variables, but kept the "percent White" so that "percent minority could be calculated by subtraction" (p. 114). The variables "percent White, percent free lunch, and number of teachers" were recoded into larger categories. They dropped the "school industrial arts" or "special education course" variables. The number of schools at risk of disclosure were then reduced to 36. They then recoded values and/or set values to missing for these specific variables for these 36 schools.

They also altered certain variables depending on their "analytic importance." These included: "number of teachers," "total school enrollment," "percent White," and "percent free lunch" because these variables are available in other public databases. The variables "grade span" and "urbanicity" were altered only if the researchers could be assured that disclosure was not at risk after making changes in the other variables. They coded "grade span" and "ethnicity" variables as "missing," instead of altering their values. Distortion was minimized by "moving schools up or down by no more than one category, after examining the distance in their scores in relation to other schools close to them" (Ingels et al., 1994b, p. 114).

Then the researchers assessed these changes to the NELS:88 data file against the Common Core Data (CCD), except no stratification by school type was required because the CCD only contains public schools. The same methodology used for the NELS:88-QED evaluation was employed with the CCD data. No schools were found to be at risk (Ingels et al., 1994b, p. 115).

[14] See QED Web site at: http://www.qeddata.com/aboutqed.htm.

[15] The code distance metric is "defined as the sum of the absolute values of the NELS/QED code differences for respective variables." There had to be at least three schools closer to the sample school for the sample school not to be considered at risk (Ingels et al., 1994b, p. 114).

Longitudinal Disclosure Risk Analysis

Disclosure risk increases as more data about an entity are collected. Ingels and colleagues state that the possibilities become more likely when the base-year and first and second follow-up data are used "in combination to identify a school" (Ingels et al.,1994b, p. 115). The authors took three measures to address the risk of disclosure in the longitudinal files (base year, first, and second follow-ups). The confidentiality edits performed for the base-year were maintained. They created "an independent set of randomized school identification number for the first follow-up schools," to make it difficult to match base year and first follow-up schools by using only the school files (p. 115). They conclude, however, that identity could "still be accomplished by analysis and deduction" (p. 115). Student records can be used to match base-year and first follow-up schools.

They then conducted an "exploratory analysis of feeder patterns on 20 first follow-up schools, utilizing only the QED school list" (p. 115). Sixteen schools were subsequently eliminated because they did not meet various criteria [e.g., "no feeder schools with at least three students or 10 or more schools closer to it than it was to itself, no feeder schools that matched themselves within the top five matches" (p. 116)]. Subsequent case study analysis on the schools led the researchers to conclude that "in no case was the signature of the first follow up/base year feeder pair so distinctive as to be absolutely unique" (p. 116).

Summary: Utility of Statistical Disclosure Limitation (SDL) Research

Statistical disclosure limitation research contributes important knowledge of the probabilities of disclosure, providing information about how data can be "transformed so that their release adequately lowers disclosure risk" (Duncan, in press, p. 1). The caveat is that there will be tradeoffs: the more severe the masking, the less useful the information; and, indeed, some techniques may "introduce bias in the inferences that are drawn" (Duncan et al., 1993, p. 147). This concern led the Ingels team to assess disclosure risk in light of the analytical potential of the data. The key lies in good decision making that depends on the scientific evaluation of risk. Assessments must be made of the disclosure risk and the constraints to be imposed on the data, so that confidentiality can be assured with a high degree of probability for the release of public use microdata files.

Survey of Organizational SDL Practices: Methodology, Data Collection, and Evaluation of Data Quality

The SDL practices survey that was conducted to provide background information for a National Research Council Committee on National Statistics workshop on confidentiality and data access. This survey focused on public use microdata sets with longitudinal data, linked administrative

data, or contextual data because these data files face special risks to deductive disclosure. A microdata file consists of records at the respondent level. Each record contains values of variables for a person, household, establishment, or other unit (U.S. Federal Committee on Statistical Methodology, 1999). "Longitudinal" or "panel" surveys are "repeated observations of the same person or other units through time" (Boruch & Pearson, 1985, p. 2). "Linked administrative data" were defined as data that could be obtained from administrative sources (e.g., Social Security Records) and linked to survey data through some unique identifier (e.g., Social Security Number). "Contextual data" are defined as characteristics (e.g., unemployment rates) of small areas (e.g., counties).

There was no practical way to construct a complete enumeration of public use microdata sets that contain longitudinal, administrative, or contextual data, nor was there any intention of generalizing to a hypothetical universe of in-scope microdata sets. Instead, the principal goal was to obtain an *understanding* of the statistical disclosure limitation (SDL) practices of organizations that have released major public use microdata sets whose sample units are potentially at risk of disclosure. Although not its primary purpose, the SDL survey also revealed the extent to which administrative procedures to provide access under controlled conditions were used by our respondents (see Massell, 1999, for a summary of administrative practices).

Various procedures were employed to gather information, including a mailout–mailback questionnaire of open-ended questions, follow-up written and telephone conversations, and Web-based searches for documents that described the both microdata files and procedures for releasing public use files. The multiple methods were necessary to clarify and verify the meaning of the information received from the respondents (see Denzin & Lincoln, 1994, on "triangulation").

Case Selection and Screening Criteria and Procedures

The first step in information gathering was to sample public use microdata sets that contained longitudinal, linked administrative, or contextual data. The sampling frame consisted of studies identified in two National Academy Press publications, *Data on America's Aging Population* (Carr, Penmarzao, & Rice, 1996) and *Longitudinal Surveys of Children* (West, Hauser, & Scanlan, 1998) that focus on currently available public use microdata for the studies of the elderly and of children, broadly defined, conducted in North America. During the information-gathering stage, several additional surveys were also identified by National Research Council staff and the survey respondents, for a total sample of 20 data files.

A contact person who was most familiar with the data set was identified within each organization through the documentation on the data file. Generally, these contact people were principal investigators or data managers. The person within the organization most familiar with issues of data confidentiality was then identified. We received information from or had communication with more than 25 people during the period of data collection. These respondents are analogous to "informants" who report on practices that are carried out in their organization in an "intrinsic" or "instrumental" (Stake, 1994) qualitative case study (on case study methodology and qualitative research in organizations more generally, see Berg, 1998; Denzin & Lincoln, 1994; Yin, 1989, 1993).

Background information on these respondents was limited to the title and unit in which the individual was located because the survey concentrated on organizational practices. Information on "title" was obtained for 23 respondents, who included project directors and managers; principal investigators; researchers, programmer, and systems analysts; directors of survey operations; survey statisticians; senior methodologists or economists; a chief of information management systems; and a vice president for research. Because follow-up conversations were necessary, we were sometimes able to obtain additional information about length of tenure in the position for some of our respondents, as well as more detailed information about the work units involved in the production of the public use files that supplemented our own knowledge about these organizations.

The data collection organizations whose microdata files met the SDL survey study criteria included government statistical agencies ($n = 3$), nonprofit firms ($n = 9$), and university-based research units ($n = 8$). These organizational units are embedded in complex institutional settings, and all have complex formal interorganizational arrangements that are a consequence of the funding of data collection as well as activities related to the production of longitudinal microdata files. These structural arrangements sometimes meant that we were referred to additional people on the project staff after our initial contacts, and they explain why the total N of microdata files is smaller than the total number of individuals with whom we spoke.

Questionnaire on SDL Organizational Practices

A questionnaire was designed to elicit information on SDL practices (see Appendix 1). Part 1 of the questionnaire requested identifying information about the respondent, organization, and microdata set. Part 2 asked the respondent a series of open-ended questions about SDL practices. This second part was subdivided into sections by type of public use data: longitudinal, linked administrative, or contextual.

Information was obtained on the units of analysis; major categories of data; some of the key variables; reference dates or periods for these variables; whether any of these variables presented special disclosure risks; what procedures were used to reduce the risks; and "any special problems" associated with persons or households that moved between waves of the survey or entered or left the sample for other reasons. Questions on administrative record linkage obtained information on units of analysis and variables included in the public use files; source(s) of the administra-

tive data; assessment of statistical disclosure risk presented by the linkage; and procedures had been applied to prevent disclosure. Questions on contextual data included a description of the geographic areas; types of variables; level of detail for these areas; precautions taken to ensure confidentiality; any other issues concerning SDL procedures; and reasons for using a particular SDL method(s).

Data Collection

In May and June 1999, a cover letter that described the purpose of the study and the questionnaire were either mailed or e-mailed to the correct contact person for each of the 16 microdata sets that had been initially selected. A reminder e-mail or phone call was sent if there were no response within a month. The one nonresponse occurred because the organization's funding was in transition, and the principal investigator reported that no one was available to answer questions about data confidentiality; however, we were able to gather limited descriptive information about this microdata set through their Web site. By the end of the data collection period, four additional data sets had been identified and information on a total of 20 microdata sets that met the study criteria had been gathered.

These 20 longitudinal microdata files contain detailed portraits of individuals, families, and households. Sample sizes range from about 5,000 to nearly 60,000 respondents. The data files are both comprehensive and sensitive. The amount of detail on individuals and households is extensive and grows significantly with every round of data collection. Several microdata sets have accumulated information on their respondents for more than 30 years. In most cases, these files contain thousands of variables. They include consumption behavior, income and earnings, health status, employment histories, eligibility for government programs, and transitions in the life course. A number of the longitudinal surveys also link survey data to administrative data. These administrative data include, for example, school records, doctor and hospital visits, and earnings, some of which derive from administrative records that are public information. Most of the microdata files contain a minimal amount of contextual/geographic data.

Information gathering through open-ended questionnaires and inside organizations is a messy business, filled with uncertainty and requiring continual evaluation of the information. In an iterative fashion, nearly always through a time-consuming and sometimes frustrating dialogue between researcher and respondents, the complexity of organizational life begins to unfold.[16] The SDL survey followed this trajectory. This process of data collection sheds light on how these organizations made decisions about releasing data and the distribution of knowledge about statistical disclosure risk.

There was considerable variation in the completeness of responses to the questionnaires. Some of the responses were detailed and clear, while others were brief, vague, or incomplete and required clarification. Sometimes questions were misunderstood. Some respondents appeared more familiar with data confidentiality terminology and concepts than others and provided us with the type of information that we expected. Those who were unfamiliar with the concepts provided answers that were not focused on the issues that the SDL survey investigated.

The scope of the information gathering process was enlarged for follow-up data collection. Data confidentiality issues that had not been part of the original questionnaire were identified, including more detailed information on the SDL techniques applied to income data; more information on restricted access data that were linked to the public use microdata sets; and more information on the sampling procedures used by organizations. Web site searches were conducted for all the microdata sets to supplement the information received through the questionnaires. The study staff recontacted respondents for more information.

Sometimes there were errors in the responses. For example, the questionnaire asked how the organization linked administrative data to their microdata sets, but respondents often supplied information about how the data user could link the administrative records to the microdata sets. It was very difficult to obtain useful answers about contextual data. Most respondents reported that their data sets contained no contextual data; this response may, however, reflect confusion over the meaning of contextual data. In at least one case, for example, our review of the data set's documentation indicated that the microdata set contained contextual data, although the respondent reported that the data set did not.[17]

Statistical Disclosure Limitation Practices Inside Organizations

The three categories of data that the SDL Organizational Practices Survey focused on present well-known disclosure risks. Nineteen of the 20 data sets identified in our study had publicly available longitudinal data. Only one organization had restricted access to its longitudinal data file, because the combination of sensitive data on the respondent's behavior and rules for participation in this survey made these data particularly vulnerable to disclosure risks. (See Appendix 2 for a description of the characteristics of and conditions for access to these microdata files.)

[16] For a discussion of the difficulties of conducting organizational surveys, see Rosenfeld, Edwards, and Thomas (1993). Yin's (1989, 1993) texts on the case study method and Zimmerman's (1970) analyses of bureaucratic life provide useful discussions about conducting case study analysis of organizational behavior.

[17] It should be noted that cognitive pretesting of the questionnaire was not conducted. It may have revealed the extent to which people had familiarity with the language and concepts of statistical disclosure limitation techniques.

How Organizations Applied Statistical Disclosure Techniques

Identifying information, such as "names," "addresses," and "social security numbers," was deleted from the public use microdata files, except in one case. In general, variables such as "age" and "income" were top-coded (see more discussion below). In many cases, linked administrative data were either suppressed, summarized at a high level, recoded, or injected with error in these public use files (see more discussion below).

In some instances, an organization restricted access to a subset of variables that could be linked to their public use microdata set because these variables were considered particularly vulnerable to confidentiality risks. About half of our respondents noted that analysis often required information that was not provided on the public use microdata file, and attempts were made to "accommodate" the research needs of analysts.[18]

Fifty percent of the organizations reported linking their survey data to administrative data. Administrative records were generally of three types: health, earnings, and school transcripts. From the few responses that described linking practices, it appears that organizations often used social security numbers to link administrative records. In the released data, the social security numbers were then suppressed and replaced with a newly created unique identifier to link administrative records to survey data. Organizations were, however, much more likely to restrict access to linked administrative data than to longitudinal data. For example, arrest history data were considered to provide various disclosure risks, and, consequently, were never publicly released.

Contextual and geographic data present similar disclosure risks because there is potential identification of specific small localities. In some cases, the two types of data were collected longitudinally, so the categories of "at-risk" data overlapped. All of the longitudinal microdata surveys included some geographic or contextual data in their public release versions; however, geographic data were generally confined to state or census division identifiers.[19] The organizations were reluctant to release detailed geographic or contextual data in the public use microdata files because of confidentiality concerns.

SDL procedures were applied to two variables that were identified as contributing to increased disclosure risks, "age" and "income" ($n = 14$ data sets). Treatment of these special "at-risk" variables varied widely by organization. There were different levels of detail at which age was available (e.g., day, month, and year of birth; month and year of birth; or year of birth sometimes given as age in years) and whether birth year was top or bottom coded. Three of the data sets top coded age (or bottom coded birth year), following guidelines proposed by statisticians at the U.S. Census Bureau.[20] Nonetheless, some organizations did not top code age, and their longitudinal panel studies now have sample persons who constitute the "very old," of more than 100 years old. Two levels of detail were reported for income (e.g., exact values of income were available or income was collapsed into categories), and top coding was applied by most organizations, although the level at which income was top coded varied widely by organization. The most strict standard was top coding the top three percent of income; the most liberal standard was top coding income above $9,999,999.

Some of the organizations reported masking other sensitive data that were particularly relevant to their microdata set, including race, ethnicity, rare health status, and educational attainment. Race groups might be collapsed into broad categories because there were few members of some racial groups in the survey or a rare mortality event might be recoded to protect these particularly sensitive health data. One respondent noted that "industry and occupation" codes "might be limited." Still another respondent commented that "marital history" data "provided a serious disclosure risk," but did not indicate why these data were unusually sensitive or whether the organization had established a policy to limit the risk.

Our discussions with respondents revealed that sometimes data were "masked" to facilitate data collection or storage, rather than because of concerns about confidentiality. For example, parental income was collected categorically in surveys administered to adolescents because adolescents might not know their parents' exact income. In some cases, age was top-coded at "99" because a two-digit field was allocated to age for data storage. An *unintended*

[18] These organizations created files with more detailed information that contained either the entire set or a subset of the original data, which might contain social security identifiers, complete address information, and variables that derive from administrative record keeping systems. One organization developed a "research file" that was a subsample of records "created in a synthetic way by swapping and combining information from different records," which could be used by analysts to test their programs and carry out preliminary analysis from their work place. The analyst's request was then submitted through a "secure" Internet server for the complete data set, the request was vetted by the organization to ensure that confidential information was not included in the output, and the output was returned as an encrypted file to the researcher. More generally, however, if restricted access to the more comprehensive set of data took place, it was through contractual agreement.

[19] The principal difference between the restricted access and public use data was that considerably more detailed contextual/geographic data appear in the restricted access file (e.g., census geocode or county identifiers, block and tract numbers). For example, some organizations released categorical county unemployment rates on their public use files and continuous county unemployment rates on their restricted access files.

[20] Based on federal agency guidelines that derive from work carried out at the U.S. Bureau of the Census, age may not be revealed for the top 3% of persons age (U.S. Federal Committee on Statistical Methodology, 1999). Thus, for example, the age of respondents in Census Bureau surveys cannot be older than 88 years at any time during the length of a panel. The "year of birth" is bottom coded to 1912, and age is recalculated for the public use files. Thus, for example, any person who is over 88 years or will age to 89+ from 1998–2000 is assigned a birth year of 1912, and will age from 84 to 88 years over the life of the panel.

consequence of these data collection and storage decisions was that data were masked.

Longitudinal data received the least attention of the "at-risk" categories of data. These data, some containing as much as 30 years of observations, were not considered to present special disclosure risks. The timing of data release led to variation in the SDL techniques applied to microdata sets. SDL standards have been modified over time, with fewer restrictions applied during the 1970s and 1980s than during the 1990s. Relatively liberal release policies for longitudinal data appeared to be, at least in part, a function of the amount of time between release of base-line and follow-ups data files. Decisions about earlier data releases did not appear to play a role in the decisions about what to release in the longitudinal files. Issues related to deductive disclosure have been brought to the attention of staff only in recent years, resulting in stricter standards for releasing public use files. The result is that older longitudinal microdata sets face special risks to deductive disclosure. Earlier data releases may contain more detailed information about respondents than would be released under current practices. Furthermore, because the data sets follow respondents over long periods of time, they contain a wealth of detailed information. The combination of changing SDL standards and the compilation of data on respondents over time may make older longitudinal data sets particularly vulnerable to disclosure. It is, however, the longitudinal data that span decades that often makes these microdata sets particularly useful to researchers.

How Respondents Assessed Disclosure Risk

As data producers, our SDL respondents were familiar with the broad issue of "data confidentiality." They were concerned about protecting the identity of respondents in the surveys they conducted and knew that direct identifiers of their respondents should not be released. The participants in our SDL study knew that name, address, and social security number were unique identifiers and had to be deleted and that income was a sensitive variable. They understood that age or date of birth needed to be masked in some way, and either deleted these and other variables in the public use files or restricted access to the data. Furthermore, nearly all the organizations did not publicly release geographic or contextual data for small areas because of concerns about data confidentiality. Similarly, linked administrative data were often confined to restricted access data.

Overall, however, the SDL respondents appeared unfamiliar with research on statistical disclosure risk or disclosure limitation methods and data confidentiality terminology. With two exceptions (see below), respondents applied simple rules for masking sensitive data in cross-sectional files. Assessment of disclosure risk on the data set was not actually carried out. Most of our respondents in university and private-sector organizations applied practical solutions, "rules of thumb," to decisions about what variables to suppress. Government agencies either followed established guidelines or deferred to a formal organizational unit responsible for making decisions about releasing public use data and applying SDL techniques; one of our respondents commented that he did not know how these officials reached a decision about which data to mask or suppress.

Only two organizations reported that they had conducted disclosure risk analysis to determine the best way to ensure the confidentiality of the respondents prior to issuing a public use file, and only one of them had performed risk analysis on longitudinal files similar to the NELS example earlier in this article. Most organizations appear to have based their SDL decisions for public use longitudinal files on the cross-sectional files. That is, with one exception, they assessed the risks of disclosure on the cross-sectional rather than longitudinal files. For example, they did not alter their policy about releasing geographic/contextual data for longitudinal data sets; in other words, all of the panels of longitudinal surveys contained the same geographic/contextual data. This could increase the risk of deductive disclosure for respondents who moved; in two cases, however, in one organization, this risk had been anticipated and geographic/contextual data were available only in the baseline public use files but omitted from later panels.

These differences across organizations appear to be real, because those who received the questionnaire had been identified by members of their own organization as the person most knowledgeable about data confidentiality issues. There appeared to be greater variation in knowledge of data confidentiality standards and disclosure risk in universities than in government agencies.[21] Those university-affiliated respondents who demonstrated the most knowledge about SDL techniques were members of survey research organizations that have long been involved with government agencies through, for example, contractual agreements for data collection, processing, or the production of public use files. These same survey research organizations had also employed statistical consultants to guide them in modern SDL practices, and practiced a "layers of access" approach, which represented protections for human subjects, for the organization, and for researchers both on and off site.

The Exigencies of Organizational Life: Interactions of Structure, Technology, and Environment

The respondents in our survey are members of large-scale organizations that are loosely integrated and operate in a highly decentralized environment, with structural units segregated one from another along functional lines of responsibility. Both the internal and external enviroment exercise enormous constraints over the introduction of new information into work practices.

[21] This assessment of the limited knowledge of disclosure risk was also confirmed by another participant in the CNStat workshop who has overseen the archiving of thousands of data sets for more than 30 years.

Structure and Technology

Control over the life cycle of data is complicated by the loose coupling of units responsible for data collection through data distribution. Responsibilities are typically partitioned according to the task to be performed: data design, collection, processing, linkage and merging, public use file creation, documentation, and distribution.

Thus, for example, instruments for collecting data were designed in offices distinct from the field units that collected the data. These two offices were also separated from data processing units responsible for preparing data for public use files and/or for analysis. Analysts to whom the data files were distributed conducted their work independently from the units that collected and processed the data files. Furthermore, these operations could be overseen by different units of the same organization or by different organizations. Their respective activities may or may not have been documented; and communication between and among the units might be rather limited, take place infrequently, or not at all.

The structurally differentiated data production life cycle also interacted with project management control, which we found differed across organizations and units. Some project managers were familiar with the nuts and bolts of data release decisions, while others were not. Rare was the situation where one principal investigator or project director was responsible for all aspects of the life cycle of data. There was also staff turnover during the lifetime of the longitudinal surveys, and several respondents reported that they were "new on the job." Herbert Simon (1991) has noted that institutional memory, in the form of memoranda and other written materials and human capital, is a critical resource for organizations. However, our survey found that historical records or project decisions and the experiences of long-tenured staff were not necessarily available to guide these new entrants.

Thus, part of the explanation for the lack of knowledge about SDL applied to longitudinal public use microdata files may reside in the difficulties associated with coordinating complex tasks in highly differentiated organizations whose units are specialized, relying on different expertise and abilities and employing different technologies to accomplish the daily work. Structural differentiation, and thus different loci of responsibility, implied different interpretations of problems and possible courses of action and, very important for managing and communicating information, of the difficult task of coordination.

Pressure from the External Environment

The social world of our respondents could be described as the "constancy of constantly changing conditions." We found that the external environment—the policy process and demands from users—had effects on the behavior of our respondents, intensifying their uncertainty and increasing the ambiguity that they experienced. The external political world was perceived as "in flux" and threatening because of the contending interests of different stakeholders in the system. Policies governing data access and confidentiality were subject to change; and the only certainty was that enacted legislation would have an impact on the functioning of their organizations. Legislative initiatives, which are responses to public concerns about confidentiality, were seen as potential threats to public use data.

Staffs were highly sensitive to the consequences of releasing data that could identify individuals. Interpretations of these data confidentiality policies influenced decisions about the release of public use microdata. One of our statistical agency respondents noted, for example, that his organization's preparation of a public use file had ceased because of a recently enacted statute, which they interpreted as preventing the distribution of public use data from their survey. Their caution was not unique, whether or not they properly interpreted the statute. What *is* important is the perceived threat from the external environment that resulted in restricting important data.

Knowledge of data user behavior also influenced decisions about release policies for longitudinal data. Staff were sensitive to the fact that data users find longitudinal data particularly useful when they contain measures of the same variables over time, and are very aware that their data would be more useful if they provided more detail in their public use microdata sets. As a result, for example, these organizations made an effort to release the same data in every panel.

One respondent made this tension clear when he stated that his organization increased data availability because of the demands of data users. The level at which income was top-coded was increased by one organization because data users complained about the lack of data on high incomes. A respondent noted that,

> We have received strong feedback from potential users regarding the importance of having state identifiers on the file. Therefore, rather than deleting state of residence altogether from the file [because a combination of state and site identifiers may have populations that are small], we have masked state of residence in the two sites where these problems exist, through use of a "data switching" procedure. This involves switching the state of residence for a portion of cases within these sites.

This respondent acknowledged that this switching introduced some measurement error for analysis at the state level or when secondary data were merged, but argued that the effect was "unlikely to have any significant effect on results at the national population analysis." Users were informed that data "switching" had taken place, but the data file documentation did not disclose the extent.

Other respondents implied that decisions to release data were based, in part, on providing as much data as possible to researchers. Public use data were "advertised" on Web sites by highlighting their detailed and longitudinal contents. This advertising is designed to increase use of these data and to demonstrate to program funders that the data

distribution function is being performed. Technological change, particularly by providing easy access to public data and statistical software through the Internet, added to these pressures to make data available to researchers. While technology facilitated the research process it also increased the risk of deductive disclosure through the availability of public data stored at Web sites.

In summary, perception of an external environment with contending interests and improved access because of technological change created uncertainties about their appropriate responses and reinforced the ambiguities and dilemmas of law and policy. Respondents had to weigh the usefulness of their data against the need for data confidentiality when they made decisions about the release of public use microdata. They thus practiced a balancing act to reconcile competing directives, "deploying strategies to cope with the clash between prescriptions" of law and policy and the unpredictability of the political environment (Brown & Duguid, 1991, p. 4).

Discussion: Locating Explanation in Theories of the Social Life of Organizations

Research in many different settings demonstrates unequivocally that organizations and their cultures, structures, and environments matter for getting the daily work done.[22] These investigations into organizations with complex technologies, conducted over more than 4 decades, and theory derived from this empirical work support our assessment that the social context of the data producing organizations constrained the search for new information by the respondents in the SDL practices survey and impeded the diffusion of innovative statistical disclosure limitation practices. The large body of empirical evidence on institutions and organizational decision making, including research on communities of practice, organizational learning, and the diffusion of innovation, yields understandings about why the organizational practices we observed do not reflect the idealized view of the "knowing organization" that uses expertise, information, and knowledge, "so that it is well informed, mentally perceptive, and enlightened" (Choo, 1998, p. xi).

The potency of structure and its interaction with culture and the environment, as well as its effects on communication processes, is understood from, for example, the early work of organizational theorists (Emory & Trist, 1965; Hall, 1972; Lawrence & Lorsch, 1967; Pugh, 1973; Selznick, 1949; Thompson, J.D., 1967a, 1967b; Thompson, V.A., 1977; Thompson & Bates, 1957; Woodward, 1958). Their studies laid the foundation for empirical research that relied on an open systems approach and brought into sharper focus the role of the environment in corporate and public sector settings and its contributions to public policy decisions and organizational learning (e.g., Alford, 1975; Feldman, 1989; Hall & Quinn, 1983; Meyer & Scott, 1983; Pheffer & Salancik, 1974, 1978; Pressman & Wildavsky, 1973; Robbin, 1984; Scott & Christenson, 1995; Scott et al., 1994; Wildavsky, 1979).

A variety of empirical investigations carried out in different settings demonstrate how organizational culture, structure, and the environment interact to create inter- and intraorganizational communication failures and limit people's search for information. The analyses of Weick (1990) and Vaughan (1996) of the Tenerife air disaster and the Challenger disaster, respectively, show that work practices contributed to horrendous technology disasters that were a consequence of inadequate or miscommunication. Robbin (1984) and Sasfy and Siegel (1981a, 1981b) found that the highly bureaucratic decision making structures of state agencies and perceptions of the external environment strongly influenced decisions that staff made to release or deny access to administrative records containing confidential information. Chatman's (1999) study shows how the walls of a state prison constrain the world view of its women inmates. Organizational hierarchies influenced how information systems were designed and later implemented (Kling & Jewett, 1995; Kling, Kraemer, Allen, Bakos, Gurbaxani, & Elliott, 1996; Napier & Smith, 2000). History and work practices were critical factors in the development of the International Classification of Disease (Bowker & Star, 1999).[23]

Evidence of the effects of organizational culture and structure on communication in the production of data also comes from work by Robbin and Frost-Kumpf (1997), who investigated two longitudinal data systems and found that the complexity of organizational, social, and technical processes introduced error that found its way into public use data files. Errors occurred because each of the units that participated in data production had its own base of authority, power, and discretion, with its own complex set of rules, procedures, and understandings about the data and how they would be collected, processed, and used. The administrative complexity of record keeping practices and the length of the communication chain for the data production process contributed to interorganizational failures in the transfer of data and information across the government agencies that these researchers examined.

Knowing and learning are dependent on the social context in which the daily work takes place, and it is this context that formed the "community of practice" of the

[22] "Culture" is broadly defined to include both collective memory and individual biography. Orr's (1990, p. 169) definition of "community memory," as she applies it to the world of her photocopier repair technicians to the larger institutional or organizational setting, is extended to "culture." "Community memory" refers to "those things known in common by all or most competent members of the technician community, the working set of current knowledge shared among technicians." Rule structures, for example, are part of institutional memory.

[23] Other research includes learning to navigate ships (Hutchins, 1990, 1991, 1996); the delivery of medical care (Cicourel, 1990); the slow response to the AIDS epidemic (Perrow & Guillen, 1990); managerial decision making in Canadian banks (Beck & Moore, 1985); and transportation and nuclear energy industrial accidents (Perrow, 2000).

respondents in our SDL survey (Chaiklin & Lave, 1996; Lave, 1988; Lave & Wenger, 1991). Thus, to understand impediments to organizational learning and the introduction of innovation into organizational practices, we need to examine "the context of the community in which the [tasks take place and tools] are used and that community's particular interpretive conventions," that is, "the practices and communities in which knowledge takes on significance-"(Brown & Duguid, 1991, pp. 12–13), because cognition is is socially situated and "progressively developed through activity" (Brown, Collins, & Duguid, 1989, p. 3).

Communities of practice create, organize, sustain, and reproduce what they know through their daily work. The organizations that our respondents worked in created the ways of communicating about their life worlds and the enterprise in which they were engaged (Brown & Duguid, 1998; Wenger, 1998). Explained theoretically in the context of failures in communicative practices in work life, Hutchins (1996, p. 52) comments that, "Lines of communication and limits on observation of the activities of others have consequences for the knowledge acquisition process." The exigencies of work and politics shaped information search and use of the research literature on SDL methods by the survey respondents, "filter[ing] out what [was] included and excluded in the calculus of decision" (Weick, 1995, p. 57), to produce that "boundly rational" worker described by Simon. In the context of their organizations, how the data production process was organized "influenced what type of information was available" (Feldman, 1989, p. 144).

The work culture of their community of practice reinforced practical ways of solving problems that ignored available information in order to reduce complexity and, at the same time, reduced the search for innovative solutions. They applied "rules of thumb" and protocols or deferred to a superior or formal rule-making, relying on rule structures, in the form of statutes, policies, regulations, and standard operating procedures that defined how to get the job done and what constituted permissible, defensible, or acceptable behavior. In other words, they created, reproduced, and legitimated existing social and political arrangements. These regulatory mechanisms reinforced the "satisficing" behavior that Simon (1997) describes.[24]

Yet it must also be recognized that there was little organizational slack to engage in information searching. Projects continually competed for attention, and there was never enough time. Not conducting a search of the literature on SDL methods or not incorporating new knowledge into the decision process for issuing public use microdata files was a rational response "when time or information-processing constraints limit to a few cues or variables the amount of information which can be processed" (Inbar, 1979, p. 17). Herbert Simon's principle of "satisficing" and his observation about the "limited information processing" capability, so fundamental to designing information systems, applied to our respondents.

Nonetheless, this explanation remains incomplete if we attribute all decision making to the constraints of rule structures and limited information processing ability. Decision making is not only a product of an organization's social history, but also of the particular biographies of its members. As we learned from the SDL survey, knowing how to apply normative rule structures for preserving data confidentiality also depended on who was on what job and for what length of time. Membership certainly carries with it "its attendant world-views," Weick (1997, p. 399) remarks, but these can be modified by personal history. Put another way, the life cycle of data reflected the life history of the organization, but this history was not independent of the biographical experiences of its members, the "practical interests, perspectives, and interpretive practices of the rule user" (Zimmerman, 1970, p. 223; see also Blumer, 1969; Garfinkel, 1967).

In summary, the work culture of our respondents provided disincentives to search for innovative solutions to the difficult problem of releasing public use longitudinal microdata files. Decision making was not a straight-forward application of the rules that governed the respondents' daily life. Institutional history, including the life cycle of data production for longitudinal microdata, introduced multiple accounts, confusion, and contradiction. Cues from the external environment were also confusing and contradictory, as our respondents noted when they discussed the effects of new legislation and the pressures to release data to researchers. As such, there remained sufficient ambiguity in the daily life of our respondents to create uncertainty, to permit discretion in decision making, to require adjudication between divergent assumptions and competing values and norms, to negotiate meaning, and to induce variation into organizational practices (March & Olsen, 1976; Robbin, 1984).

Coda on Theoretical Explanation and Policy Implications

Our project began as a small, exploratory survey to determine what statistical disclosure limitation methods were applied by organizations that produced public use microdata files containing longitudinal, administrative, and contextual data. The survey originated out of concern that "an effective governmental statistical system," which depends on "data that society collects and maintains about itself, its institutions and its citizens" (Duncan & Pearson, 1991b, pp. 237–238) requires a knowledge base about how

[24] "Satisficing" refers to how alternatives are evaluated before a decision is made. Simon's seminal article, "A Behavioral Model of Rational Choice" (1955), argued that the conception of "Rational Man" was not an accurate description of how people made decisions; indeed, it was "incompatible with the access to information and computational capacities that are actually possessed by organisms, including man . . ." (p. 7) People do not have a "well-organized and stable system of preferences" (p. 10), or are they able to evaluate all alternatives and reach the optimal solution. Rather, in "complex choice situations," they simplify in order to make a choice to "find an approximate model of manageable proportions" (p. 10) and choose the first satisfactory alternative; in other words, they "satisfice."

well statistical solutions are being applied to reduce disclosure risk.

At the outset of our project, the staff responsible for the SDL survey assumed that the corpus of research and information about good practices had been communicated to organizations responsible for the production of public use longitudinal microdata files. Research into disclosure risk had been conducted for more than 25 years, and many journal articles had been published on the subject. Statistical agencies had also offered guidelines for good practice that are circulated throughout the federal government and published at government web sites. Informational materials and workshops are also available through professional association like the American Statistical Association (American Statistical Association, 1999a, 1999b). Yet, the responses to the SDL survey questionnaire and the follow-up conversations revealed that many of the respondents did not know the language and application of SDL methods.

The findings concerning SDL practices were unexpected and subsequently propelled a search for explanation. An intensive review of the original responses and the follow-up interviews with our respondents provided insights into a host of factors that appear to have contributed to their lack of familiarity with statistical disclosure risk for longitudinal microdata files. The practices that the survey respondents engaged in were a consequence of common-sense judgments, tacit, taken-for-granted understandings, practical reasoning, and the routinized procedures of every day work life. Through their accounting practices, respondents made visible the social order of institutions and organizations, "an understanding of the lived experience" (Holstein & Gubrium, 1994, p. 262). History and politics left their imprints on these organizations and on the practice of producing public use files. The SDL work setting was itself a community of practice whose members experienced failures in communication and information exchange.

Redefining the findings from the SDL survey more generally as a problem of decision making in large-scale, complex organizations locates explanation for impediments to organizational learning and communication of new knowledge as a function of structural and technological complexity and the interactions between organizations and their environments. Redefining work life inside these organizations as a "community of practice" also provides insights into the problems of translating the research of statisticians into the practices of a community of nonstatisticians who are engaged in the release of public use longitudinal microdata files and into the policy solutions that flowed from our discovery.

We discovered that the properties of organizations, including their structures, environments, and meaning systems, contributed to or impeded knowledge use about disclosure risk and SDL methods associated with longitudinal microdata. These properties had significant effects on information production, coordination, and communication, as well as on the acquisition of new knowledge from outside the work unit and organization.

The theoretical lens of "communities of practice" illuminates our understanding of the seeming deficits in the knowledge (and therefore language) of SDL concepts and terminology of our respondents, and also makes explicit why they ignored a large formal literature on statistical disclosure limitation methods. Employing a more nuanced language of explanation, Brown and Duguid (1991, p. 12) suggest that we understand the respondents' SDL practices as a consequence of "a wide range of materials that include ambient social and physical circumstances and the histories and social relations of the people involved."

The policy problem is how to ensure that already available knowledge of the community of practice of statisticians is incorporated in the corpus of knowledge of the community of practice of data producers and distributors and how to create effective communication channels for knowledge use to take place. The policy solution is to improve and institutionalize the communication of information on how to appraise and treat statistical disclosure risk and to make data managers more competent and effective.

The "community of practice" theoretical perspective also offers a policy solution for actively engaging the participants in microdata public use file production in a search for new knowledge—as in Lave's (1998, p. 6) articulation of the nature of social relations and context: "Participation in everyday life may be thought of as a process of changing understanding in practice, that is, as learning." Social participation is constitutive of learning and knowing. Extending Lave's and others' conceptions of communities of practice to conceptualize organizations as "social networks of learning" (see, for example, Pesconsolido & Rubin, 2000; Powell, Koput, & Smith-Doerr, 1996) also yields a policy recommendation to engage the various types of expertise needed for assessing disclosure risk and applying appropriate solutions. That very old notion of the role of a "boundary spanner" can also be usefully resurrected (Allen, 1970; for recent treatments that situate boundary spanning in the communities of practice school, see Albrechtsen & Jacob, 1998; Davenport & Hall, in press; Jacob & Albrechtsen, 1997). Boundary spanners can potentially solve the problem of heterogeneous communities of practice—social worlds each with their own linguistic traditions that impede communication—and help construct a "common ground" for shared communicative environments (for achieving "common ground" see Clark, 1985; Krauss & Fussell, 1991).

Conceptualized as choice between competing values and contending interests of stakeholders, the problem for our SDL respondents is: how to adjudicate, balance, and reconcile abstract, ambiguous, and contradictory public policies on privacy, confidentiality, and records release to carry out policy and social research in an unpredictable world. One consequence of ambiguity and uncertainty in this terrain of action is, however, that there are usually multiple repertoires available for negotiating, reconciling, or achieving provisional consensus about the problematic status of these rule structures (see March & Olsen, 1989, pp. 21–22). Ambiguity conceived as opportunity implies that the terrain

can also be a source of innovation (Kruskal, 1981, p. 513). Critical, then, for our respondents is deciding which cues to attend to, to help them make sense of the situation and to be "inventive [and] 'ignore precedent, rules, and traditional expectations' and break conventional boundaries" (Brown & Duguid, 1991, p. 17, also quoting Orr, 1990).

Conflicts in the "ground rules" will always engender cross pressures that are unavoidable. The complexity of this political environment will result in administrative solutions that deny or restrict access to data for social research. But these cross pressures are not necessarily a liability. They can lead to creative solutions that utilize a knowledge base of substantial research on statistical disclosure risk and statistical disclosure limitation methods, so that both data access and data protection remain privileged values. In other words, cross pressures are also opportunities for organizational learning. The work setting does not foreclose improvisation or a search for solutions to reconcile the competing values. Indeed, developing administrative procedures for conditions of restricted access can be viewed, along with other practices, as innovative solutions to reduce risk and the dilemmas that our respondents faced.

Personal information is the lifeblood of organizations. It is a knowledge asset, extraordinarily valuable for the conduct of government, corporate, and scientific life. Governments require personal information to administer programs, provide services, and ensure accountability. Business firms rely on personal information to manage transactions with customers, conduct research to increase market share, and identify potential customers of their products. Schools and health care providers rely on personal information on students and patients to assess progress and provide appropriate services. Research on human subjects and businesses, as well as many other entities, often requires access to individually identifiable information.

Nonetheless, information privacy has become a very important and high-profile policy issue because of the very large amount of personal information collected by public and private sector organizations. The public debate is, however, principally about how personal information will be *used*. In the end, the contentious debate about the intrusiveness of the 2000 decennial census subsided when citizens understood that their information was a critical information resource for distributing federal funds to their communities (Lott, 2000). Moreover, although public opinion surveys indicate that people exhibit anxiety about information privacy, their concerns reflect a lack of confidence in the private sector's promises to safeguard personal information. They remain very enthusiastic about the benefits of information technology (Pew Research Center, 1999) and cautiously willing to share information on-line (McGuire, 2000), recognizing the utility of a transaction that exchanges personal information for desired services.

The scientific discourse about personal privacy that revolves around the production and release of statistical data is a subset of this larger policy debate about information privacy, one that addresses the cross pressures of maintaining confidentiality and ensuring access. As with all other public policy issues, there are important moral and ethical issues that figure in the calculus of and trade-offs between efficiency and effectiveness, and these cannot be ignored. These contradictory impulses are critical to understand for their policy implications, and the scientific community must be sensitized and respond to them in appropriate ways. Although the technical issue of public use microdata file release is far less visible, it, too, warrants public discussion.

Acknowledgments

We thank Ralph Brower, Blaise Cronin, Elizabeth Davenport, George Duncan, Elin Jacob, Kathleen de la Pena McCook, and anonymous reviewers for their very valuable assistance in preparing this article.

References

Alford, R.R. (1975). Health care politics: Ideological and interest group barriers to reform. Chicago: The University of Chicago Press.

Albrechtsen, H., & Jacob, E. (1998). The dynamics of classification systems as boundary objects for cooperation in the electronic library. Library Trends, 47(2), 293–312.

Allen, T.J. (1970). Roles in technical communication networks. In Neson, C.E. & Pollock, D.K. (Eds.), Communication among scientists and engineers (pp. 191–208). Lexington, MA: Heath Lexington Books.

American Association of Public Opinion Research. (1999). Best practices for survey and public opinion research. Ann Arbor, MI: Author. Retrieved October 5, 2000 from the World Wide Web: http://www.aapor-.org/ethics/best.html#best11.

American Society for Information Science. (1997, February/March). 1997 ASIS mid-year meeting. ASIS heads west for a look at information privacy, integrity and data security. Bulletin, 23(3), 2–3.

American Statistical Association, Committee on Privacy and Confidentiality. (1999a). Information about statistical disclosure limitation methods. Retrieved on April 17, 2000 from the World Wide Web: http://www.erols.com/dewolf/protect/sdlnfo.htm.

American Statistical Association. Committee on Privacy and Confidentiality. (1999b). Procedures for restricted data access. Retrieved on April 17, 2000 from the World Wide Web: http://www.erols.com/dewolf/protect/raccess.htm.

Americans facing loss of personal privacy. (2000, 3 March). Congressional Record, H1288. Retrieved June 6, 2000 from the World Wide Web: http://thomas.loc.gov/.

Beck, B.E.F., & Moore, L. (1985). Linking the host culture to organizational variables. In Frost, P.J., Moore, L.F., Louis, M.R., Lundberg, C.C., & Martin, J. (Eds.), Organizational culture (pp. 335–354). Thousand Oaks, CA: Sage Publications, Inc.

Bennett, C.J. (1997). Arguments for the standardization of privacy protection policy: Canadian initiatives and American and international responses. Government Information Quarterly, 14(4), 351–362.

Bennett, C.J. (1998). Convergence revisited: Toward a global policy for the protection of personal data? In Agre, P.E. & Rotenberg, M. (Eds.), Technology and privacy: The new landscape (pp. 98–123). Cambridge, MA: The MIT Press.

Bennett, C. J., & Raab, C.D. (1997). The adequacy of privacy: The European Union data protection directive and the North American response. The Information Society, 13(3), 245–263.

Berg, B.L. (1998). Qualitative research methods for the social sciences (3rd ed.). Boston: Allyn and Bacon.

Bethlehem, J.A., Keller, W.J., & Pannekok, J. (1989). Disclosure control of microdata. Journal of the American Statistical Association, 85, 38–45 (cited in Winkler, 1997).

Blumer H. (1969). Symbolic interactionism: Perspective and method. Englewood Cliffs, NJ: Prentice-Hall.

Boruch, R.F., & Pearson, R.W. (1985). The comparative evaluation of longitudinal surveys New York: Social Science Research Council.

Bowker, G., & Star, S.L. (1999). Sorting things out: Classification and its consequences. Boston: The MIT Press.

Brown, J.S., & Duguid, P. (1991). Organizational learning and communities-of-practice. Palo Alto, CA: Xerox Parc. Retrieved October 27, 2000 from the World Wide Web: http://www.parc.xerox.com/ops/members/brown/papers/orglearning.html (This article first appeared in Organization Science, February 1991.)

Brown, J.S., & Duguid, P. (1998). Organizing knowledge. Palo Alto, CA: Xerox Parc. Retrieved October 27, 2000 from the World Wide Web: http://slofi.com/oganizi.htm (This article first appeared in 1998, California Management Review, 40(1) 90–111.)

Brown, J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Palo Alto, CA: Xerox Parc. Retrieved October 27, 2000 from the World Wide Web: http://www.slofi.cm/situated.htm (This article was published in 1989, Educational Researcher, January–February, 32–42.)

Carr, D., Penmarazo, A., & Rice, D.P. (Eds.). (1996). Improving data on America's aging population. Summary of a workshop. Washington, DC: National Academy Press.

Caruso, D. (1999, August 30). Consumers' desire for information privacy ignored. The New York Times. Retrieved June 8, 2000 from the World Wide Web: http://www.nytimes.com/library/tech/99/08/biztech/articles/30digi.html.

Cate, F.H. (1997). Privacy in the information age. Washington, DC: Brookings Institution Press.

Chaiklin, S., & Lave, J. (Eds.). (1996). Understanding practice: Perspectives on activity and context. New York: Cambridge University Press.

Chatman, E.A. (1999). A theory of life in the round. Journal of the American Society for Information Science, 50(3), 207–217.

Choo, C.W. (1998). The knowing organization: How organizations use information to construct meaning, create knowledge, and make decisions. New York: Oxford University Press.

Cicourel, A.V. (1990). The integration of distributed knowledge in collaborative medical diagnosis. In Galegher, J., Kraut, R.E., & Egido, C. (Eds.), Intellectual teamwork: Social and technological foundations of cooperative work (pp. 221–242). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Clark, H.H. (1985). Language use and language users. In Lindzey, G., & Aronson, E. (Eds.), Handbook of social psychology (pp. 179–231). New York: Random House.

Clausing, J. (2000, February 7). Report rings alarm bells about privacy on the Internet. The New York Times. Retrieved February 7, 2000 from the World Wide Web: http://www.nytimes.com/library/tech/00/04/biztech/articles/07priv.htm.

Clement, A. (1994). Computing at work: Empowering action by low-level users. Communications of the ACM, 37(1), 52–65.

Cox, L.H. (1991, August). [Enhancing access to microdata while protecting confidentiality]: Comment. Statistical Science, 6(3), 232–234.

Davenport, E., & Hall, H. (In press). Organizational knowledge and communities of practice.

Denzin, N.K., & Lincoln, Y.S. (Eds.). (1994). Handbook of qualitative research. Thousand Oaks, CA: Sage Publications, Inc.

Duncan, G.T. (in press). Confidentiality and statistical disclosure limitation. International Encyclopedia of the Social and Behavioral Sciences.

Duncan, G.T., Jabine, T.B., & de Wolf, V.A. (Eds.). (1993). Private lives and public policies. Washington, DC: National Academy Press.

Duncan, G.T., & Pearson, R. (1991a, August). Enhancing access to microdata while protecting confidentiality: Prospects for the future. Statistical Science, 6(3), 219–232.

Duncan, G.T., & Pearson, R. (1991b, August). [Enhancing access to microdata while protecting confidentiality: Prospects for the future] Rejoinder. Statistical Science, 6(3), 237–239.

Emory, F.E., & Trist, E.L. (1965). The causal texture of organizational environments. Human Relations 18(1), 21–32.

European Commission. Data Protection Working Party. (1997). Data protection: Annual report of the data protection working party. First annual report. Brussels: The European Commission. Retrieved November 6, 2000 from the World Wide Web: http://europa.eu.int/comm/internal_market/en/media/dataprot/wpdocs/wp3en.htm.

Feldman, M.S. (1989). Order without design: Information production and policy making. Stanford, CA: Stanford University Press.

Fienberg, S.E., & Willenborg, L.C.R.J. (Eds.). (1998, December). Special issue on disclosure limitation methods for protecting the confidentiality of statistical data. Journal of Official Statistics, 14(4).

Fienberg, S.E., Martin, M.E., & Straf, M.E. (Eds.). (1985). Sharing research data. Washington, DC: National Academy Press.

Frishberg, M. (2000, April 20). U.S. confused about privacy. Wired News. Retrieved May 2, 2000 from the World Wide Web: http://www.wired.com/news/politics/0,1283,35979,00.html.

Garfinkel, H. (1967). Studies in ethnomethodology. Englewood Cliffs, NJ: Prentice-Hall.

Hall, R.H. (1972). Organizations: Structure and process. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Hall, R.H., & Quinn, R.E. (Eds.). (1983). Organizational theory and public policy. Thousand Oaks, CA: Sage Publications, Inc.

Holstein, J.A., & Gubrium, J.F. (1994). Phenomenology, ethnomethodology, and interpretive practice. In Denzin, N.K. & Lincoln, Y.S. (Eds.), Handbook of qualitative research (pp. 262–272). Thousand Oaks, CA: Sage Publications, Inc.

Hutchins, E. (1990). The technology of team navigation. In Galegher, J., Kraut, R.E., & Egido, C. (Eds.), Intellectual teamwork: Social and technological foundations of cooperative work (pp. 191–220). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hutchins, E. (1991). The social organization of distributed cognition. In Resnick, L.B., Levine, J.M., & Teasley, S.D. (Eds.), Perspectives on social shared cognition (pp. 283–307). Washington, DC: American Psychological Association.

Hutchins, E. (1996). Learning to navigate. In S. Chaiklin & J. Lave (Eds.), Understanding practice: Perspectives on activity and context (pp. 35–63). New York: Cambridge University Press.

Inbar, M. (1979). Routine decision-making: The future of bureaucracy. Thousand Oaks, CA: Sage Publications, Inc.

Ingels, S.J., Dowd, K.L., Baldride, J.D., Stipe, J.L., Bartot, V.H., & Frankel, M.R. (1994a, September). National education longitudinal study of 1988. Second follow-up: Student component data file user's manual. NCES Technical Report 94-374. Washington, DC: National Center for Education Statistics.

Ingels, S.J., Scott, L.A., Rock, D.A., Pollack, J.M., & Rasinski, K.A. (1994b, October). National education longitudinal study of 1988. First follow-up final technical report: Base year to first follow-up. NCES Technical Report 94-632. Washington, DC: National Center for Education Statistics.

Jabine, T. B. (1993). Procedures for restricted data access. Journal of Official Statistics, 9(2), 537–589.

Jacob, E.K., & Albrechtsen, H. (1997). Constructing reality: The role of dialogue in the development of classificatory structures. In McIlwaine, I.C. (Ed.), Knowledge organization for information retrieval. Proceedings of the 6th International Study Conference on Classification Research, 14–16 June 1997, London (pp. 42–50). The Hague: International Federation for Documentation.

Kennickell, A. (1999a, October). Data simulation and disclosure limitation in the Survey of Consumer Finances. Paper presented at the Workshop on Confidentiality of and Access to Research Data Files of the Committee on National Statistics, National Research Council, Washington, DC.

Kennickell, A. (1999b, October). Multiple imputation in the Survey of Consumer Finances. Paper presented at the Workshop on Confidentiality of and Access to Research Data Files of the Committee on National Statistics, National Research Council, Washington, DC.

Kim, J.J., & Winkler, W.E. (1995). Masking microdata files. Proceedings of the Section on Survey Research Methods of the American Statistical

Association (pp. 114–119). Arlington, VA: American Statistical Association.
Kling, R., & Jewett, T. (1995, March). The social design of worklife with computers and networks: An open natural systems perspective (distribution draft). Retrieved October 31, 2000 from the World Wide Web: http://www.slis.indiana.edu/kling/pubs/worknt.html.
Kling, R., Kraemer, K.L., Allen, J.P., Bakos,Y. Gurbaxani, V., & Elliott, M. (1996, February). Transforming coordination: The promise and problems of information technology in coordination. Retrieved October 31, 2000 from the World Wide Web: http://www.slis.indiana.edu/kling/pubs/CTCT97B.htm
Kraus, R.M., & Fusell, S.R. (1991). Constructing shared communicative environments. In Resnick, L.B., Levine, J.M., & Teasley, S.D. (Eds.), Perspectives in socially shared cognition (pp. 172–202). Washington, DC: American Psychological Association.
Kruskal, W. (1981). Statistics in society: Problems unsolved and unformulated. Journal of the American Statistical Association, 76(375), 505–515.
Lave, J. (1988). Cognition in practice: Mind, mathematics, and culture in everyday life. New York: Cambridge University Press.
Lave, J. (1998). The practice of learning. In Chaiklin, S. & Lave, J. (Eds.), Understanding practice: Perspectives on activity and context (pp. 3–34). New York: Cambridge University Press.
Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. New York: Cambridge University Press.
Lawrence, P.R., & Lorsch, J.W. (1967). Organizations and environment. Boston: Harvard University Press.
Lott, T. (2000, April 6). Census 2000 is Mississippi's future (press release). Retrieved July 20, 2000, from the World Wide Web: http://lott.senate.gov/news/2000/0406.census.html
Mackie, C., & Bradburn, N. (Eds.). (2000). Improving access to and confidentiality of research data: Report of a workshop. Washington, DC: National Academy Press.
March, J.G., & Olsen, J.P. (1976). Ambiguity and choice in organizations. Bergen, Norway: Universitetsforlaget.
March, J.G., & Olsen, J.P. (1989). Rediscovering institutions: The organizational basis of politics. New York: The Free Press.
Martin, E. (2000). Public opinion changes during two censuses. Paper presented at the Decennial Census Advisory Committee, September 21, 2000, Washington, DC.
Martin, S.B. (1998). Information technology, employment, and the information sector: Trends in information employment 1970–1995. Journal of the American Society for Information Science, 49(10), 1053–1069.
Martin, M.E., & Straf, M.E. (Eds.). (1985). Sharing research data. Washington, DC: National Academy Press.
Massell, P. (1999). Review of data licensing agreements of U.S. government agencies and research organizations. Paper presented at the Workshop on Confidentiality of and Access to Research Data Files of the Committee on National Statistics, National Research Council, Washington, DC, October 14–15, 1999.
McAdams, R., Smelser, N.J., & Treiman, D.J. (Eds.). (1982). Behavioral and social science research: A national resource. Part I. Washington, DC: National Academy Press.
McCullagh, D., & Morehead, N. (2000, July 20). FTC goes public with privacy. Wired News. Retrieved July 25, 2000 from the World Wide Web: http://www.wired.com/news/print/0,1294,37695,00.html
McGuire, D. (2000, November 30). Americans cautiously willing to share info online study. Newsbyte.com. Retrieved November 30, 2000, from the World Wide Web: http://www.newsbytes.com/news/00/15801.html
McMillen, M. (1999, October). National Center for Education Statistics: Data licensing agreements. Paper presented at the Workshop on Confidentiality of and Access to Research Data Files of the Committee on National Statistics, National Research Council, Washington, DC.
Meyer, J.W., & Scott, W.R. (1983). Organizational environments: Ritual and rationality. Thousand Oaks, CA: Sage Publications, Inc.
Mosaquera, M. (2000, June 15). Lawmakers seek balance in privacy legislation. The New York Times. Retrieved June 15, 2000 from the World Wide Web: htpp://www.ntimes.com/library/. . ./TWB20000615S0017.html
Napier, M.E., & Smith, K.A. (2000, Spring). Earth's largest library—Panacea or anathema? A socio-technical analysis. Center for Social Information Working Paper. Bloomington, IN: School of Library and Information Science. Retrieved October 31, 2000 from the World Wide Web: http://www.slis.indiana.edu/CSI/wp00-02.html
National Research Council. Committee on National Statistics. (1979). Privacy and confidentiality as factors in survey response. Washington, DC: National Academy of Science.
National Science Board, Special Commission on the Social Sciences. (1969). Knowledge into action: Improving the nation's use of the social sciences. Washington, DC: National Science Foundation.
Nie, N.H., & Junn, J. (2000, May 4). America's experience with Census 2000: A preliminary report. Retrieved May 24, 2000, from the World Wide Web: http://www.intersurvey.cm/about_intersurvey/press/05042000_census.htm
Nie, N., Junn, J., & Slotwiner, D. (2000, May). The 2000 census civic mobilization effort: Influences on participation. Paper presented at the annual meeting of the American Association of Public Opinion Research, May 18–21, 2000, Portland, OR. Available from the author Junn: aaporcensus2.ppt.
Norman, J. (2000, March 27). From pep rallies to pet peeves, census 2000 gets most hype, flap. Milwaukee Journal Sentinel, p. 1B.
Norwood, J.L. (1991, August). [Enhancing access to microdata while protecting confidentiality: prospects for the future]: Comment. Statistical Science, 6(3), 236–237.
Okie, S. (2000, April 16). Groups warn of breaches in privacy laws for patients. The Washington Post. Retrieved June 8, 2000 from Lexis-Nexis Academic Universe from the World Wide Web: http://web.lexis-nexis.com/universe.
Orr, J. (1990). Sharing knowledge, celebrating identity: War stories and community memory in a service culture. In Middleton, D.S. & Edwards, D. (Eds.), Collective remembering: Memory in society (pp. 169–189). Beverley Hills, CA: Sage Publications.
Owings, J. (2000). NELS:88/2000 fourth follow-up: An overview. Washington, DC. National Center for Education Statistics. Retrieved November 1, 2000, from the World Wide Web: http://nces.ed.gov/2000301.pdf
Perrine, K. (2000, November 1). TRUSTe to launch EU safe harbor seal. TheStandard. Retrieved November 6, 2000 from the World Wide Web: http://thestandard.com/article/article_print/0.1153,19846,00.html
Perrow, C. (2000). Normal accidents (2nd ed.). Princeton, NJ: Princeton University Press.
Perrow, C., & Guillen, M.F. (1990). The AIDS disaster: The failure of organizations in New York and the nation. New Haven, CT: Yale University Press.
Pew Research Center for The People & The Press. (1999, July 3). Public perspectives on the American century: Technology triumphs, morality falters. Washington, DC: Pew Research Center. Retrieved July 19, 2000, from the World Wide Web: http://wwww.people-press.org/mill1rpt.htm
Pheffer, J., Salancik, G.R. (1974). Organizational decision making as a political process: The case of the university budget. Administrative Science Quarterly, 19, 135–151.
Pheffer, J., & Salancik, G.R. (1978). The external control of organizations: A resource dependence perspective. New York: Harper and Row Publishers, Inc.
Powell, W.W., Koput, K.W., & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. Administrative Science Quarterly, 41, 116–145.
Pressman, J.L., & Wildavsky, A. (1973). Implementation. Berkeley, CA: University of California Press.
Pugh, D.S. (1973). The measurement of organization structures: Does context determine form? Organizational Dynamics, Spring, 19–34.
Robbin, A. (1984). A phenomenology of decisionmaking: Implementing information policy in state health and welfare agencies. Unpublished doctoral dissertation, University of Wisconsin–Madison.
Robbin, A. (2001a). The loss of personal privacy and its consequences for social research. Journal of Government Information.
Robbin, A. (2001b). Interpretations of privacy and confidentiality rules by government agencies. Paper delivered at the annual meeting of the Population Association of America, March 30, 2001, Washington, DC.

Robbin, A., & Frost-Kumpf, L. (1997). Extending theory for user-centered information services: Diagnosing and learning from error in complex statistical data. Journal of the American Society for Information Science, 48(2), 96–121.

Robbin, A., & Koball, H. (2000). Statistical disclosure limitation (SDL) practices of organizations that distribute public use microdata. Paper presented at the annual meeting of the American Association of Public Opinion Research, Portland, OR, May 18–21, 2000.

Robbin, A., Koball, H., & Jabine, T. (1999). A survey of statistical disclosure limitation (SDL) practices of organizations that distribute public use microdata. Paper presented at the Workshop on Confidentiality of and Access to Research Data Files of the Committee on National Statistics, National Research Council, Washington, DC, October 14–15, 1999.

Rosenfeld, P., Edwards, J.E., & Thomas, M.D. (Eds.). (1993, March/April). Improving organizational surveys: New directions and methods. American Behavioral Scientist, 36(4). (Special issue devoted to conducting organizational surveys.)

Rosenbaum, D.E. (2000, April 1). Seeking answers, census is stirring privacy questions. The New York Times. Retrieved April 1, 2000 from the World Wide Web: http://www.nytimes.com/library/national/040100privacy-census.html

Sasfy, J.H., & Siegel, L.G. (1981a). A study of research access to confidential criminal justice agency data. McLean, VA: The MITRE Corporation.

Sasfy, J.H., & Siegel, L.G. (1981b). The impact of privacy and confidentiality laws on research and statistical activity. McLean, VA: The MITRE Corporation.

Scott, W.R., & Christensen, S. (Eds.). (1995). The institutional construction of organizationals. Thousand Oaks, CA: Sage Publications, Inc.

Scott, W.R., Meyer, J.W., & Associates. (1994). Institutional environments and organizations: Structural complexity and individualism. Thousand Oaks, CA: Sage Publications, Inc.

Selznick, P. (1949). TVA and the grass roots. Berkeley: University of California Press.

Simon, H.A. (1979 [1955]). A behavioral model of rational choice. Quarterly Journal of Economics, 69, 99–118. Reprinted 1979, in Models of Thought (pp. 7–19). New Haven, CT: Yale University Press.

Simon, H.A. (1991). Bounded rationality and organizational learning. Organization Science, 2(1), 125–134.

Simon, H.A. (1997). Administrative behavior: A study of decision-making processes in administrative organizations (4th ed.). New York: The Free Press.

Stake, R.S. (1994). Case studies. In Denzin, N.K. & Lincoln, Y.S. (Eds.), Handbook of qualitative methods (pp. 236–247). Thousand Oaks, CA: Sage Publications, Inc.

Star, S.L. (1991). Invisible work and silenced dialogues in knowledge representation. In Erikssson, I., Kitchenham, B., & Tijdens, K. (Eds.), Women, work and computerization (pp. 81–92). Amsterdam: North-Holland.

Suchman, L. (1996). Supporting articulation work. In Kling, R. (Ed.), Computerization and controversy: Value conflicts and social choices (2nd ed.) (pp. 407–423). New York: Academic Press.

Sullivan, C. (1992, September 22). An overview of disclosure principles. Research Report Series no. 92-09. Washington, DC: U.S. Bureau of the Census. Retrieved August 30, 2000 from the World Wide Web: http://www.census.gov/srd/www/rr92-9.pdf

Task Force on the 2000 PUMS. (May 2000). Census 2000 PUMS Report. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Retrieved July 21, 2000 from the World Wide Web: http://www.ipums.umn.edu/~census2000/background.html.

Thompson, J.D. (1967a). Organizations in action. New York: McGraw-Hill Book Company.

Thompson, J.D. (1967b). Technology, organization, and administration. Administrative Science Quarterly, 2(3) (December), 325–343.

Thompson, J.D., & Bates, F.L. (1957). Technology, organization, and administration. Administrative Science Quarterly, 2(3), 325–343.

Thompson, V.A. (1977). Modern organization (2nd ed.). University, AL: The University of Alabama Press.

U.S. Department of Commerce. (1999, November 8). U.S. Secretary of Commerce William M. Daley calls for consumer privacy protection in online profiling. (News release). Retrieved June 6, 2000 from the World Wide Web: www.doc.gov/20release.html

U.S. Department of Health and Human Services, Health Care Financing Administration. (1992). Notices of proposed name changes and additional routine uses for existing system of records. Federal Register 57, no. 219 (12 November 1992): 53763.

U.S. Federal Committee on Statistical Methodology. (1978). Report on statistical disclosure-avoidance techniques. Statistical Policy Working Paper 2. Washington, DC: U.S. Department of Commerce.

U.S. Federal Committee on Statistical Methodology. (1994). Report on statistical disclosure limitation methodology. Statistical Policy Working Papers 22. Washington, DC: U.S. Department of Commerce.

U.S. Federal Committee on Statistical Methodology. (1999, July). Checklist on disclosure potential of proposed data releases. Washington, DC: U.S. Office of Management and Budget.

U.S. Federal Trade Commission. (1998, June). Privacy online: A report to Congress. Retrieved April 24, 2000 from the World Wide Web: http://www.ftc.gov/reports/privacy/toc.htm

U.S. Federal Trade Commission. (2000, May). Privacy online: Fair information practices in the electronic marketplace: A report to Congress. Washington, DC: FTC. Retrieved June 6, 2000 from the World Wide Web: http://www.ftc.gov/ as a pdf file.

U.S. House Committee on Government Operations, Legislation and National Security. (1982). Federal government statistics and statistical policy: Hearing before a Subcommittee of the Committee on Government Operations, 97th Cong., 2nd sess., 3 June 1982.

U.S. Privacy Protection Study Commission. (1977). Personal privacy in an information society. Washington, DC: U.S. Government Printing Office.

Vaughan, D. (1996). The Challenger launch decision: Risky technology, culture, and deviance at NASA. Chicago: The University of Chicago Press.

Weick, K. (1990). The vulnerable system: An analysis of the Tenerife air disaster. Journal of Management, 16(3), 571–593.

Weick, K.E. (1995). Sensemaking in organizations. Thousand Oaks, CA: Sage Publications, Inc.

Weick, K. (1997). Book review symposium: The [Challenger launch decision:] Risky technology, culture, and deviance at NASA. Administrative Science Quarterly, 42(2), 395–402.

Weiss, C.H. (1977a). Research for policy's sake: The enlightenment function of social research. Policy Analysis 3(4), 531–545.

Weiss, C.H. (1977b). Using social research in public policy making. Lexington, MA: Lexington Books.

Wenger, E. (1998). Communities of practice: Learning, meaning, and identity. New York: Cambridge University Press.

West, K.K., Hauser, R.M., & Scanlan, T.M. (Eds.). (1998). Longitudinal surveys of children. Washington, DC: National Academy Press.

Westin, A.F. (Ed.). (1971). Information technology in a democracy. Cambridge, MA: Harvard University Press.

Westin, A.F. (1976a). Computers, health records, and citizen's rights. Washington, DC: U.S. Department of Commerce.

Westin, A.F. (1976b). Privacy and freedom. New York: Atheneum.

Westin, A.F., & Baker, M.A. (1972). Databanks in a free society: Computers, record-keeping and privacy. New York: Quadrangle/The New York Times Book Co.

Wildavsky, A. (1979). The politics of the budgetary process (3rd ed.). Boston: Little, Brown and Company.

Winkler, W.E. (1997). Views on the production and use of confidential microdata. Research Report Series no. 97-1. Washington, DC: U.S. Bureau of the Census, 1997. Retrieved August 30, 2000 from the World Wide Web: http://www.census.gov/srd/www/rr97-1.pdf

Winkler, W.E. (1998). Producing public-use microdata that are analytically valid and confidential. Research Report Series no. 98-02. Washington,

DC: U.S. Bureau of the Census, 1998. Retrieved June 7, 2000 from the World Wide Web: http://www.census/gov/srd/www/rr98-2.pdf
Woodward, J. (1984 [1958]). Management and technology. Reprinted in Pugh, D.S. (Ed.), Organization theory, 2nd ed. (pp. 52–66). New York: Viking Penguin Inc.
Yin, R.K. (1989). Case study research: Design and methods (rev. ed.). Newbury Park, CA: Sage Publications, Inc.
Yin, R.K. (1993). Applications of case study research. Newbury Park, CA: Sage Publications, Inc.
Zimmerman, D. (1970). The practicalities of rule use. In J. Douglas (Ed.), Understanding everyday life (pp. 221–238). Chicago: Aldine.

Appendix 1. SDL Organizational Practices Survey Questionnaire

[Note that this is a facsimile of the questionnaire. In the original questionnaire, the font size is at 12 point and space was left between each question.]

Workshop on Confidentiality of and Access to Research Data Files
National Academy of Sciences

Instructions

Please record your answers in the format most convenient for you: on this form, in an attachment if responding by e-mail, or on separate sheets. Use the question numbers to identify answers. To save your time, if there is documentation available that answers some questions, please include it with your response and refer to the relevant documents in your answers to the questions.

A. Identification

1. Name of survey:
2. Agency or organization:
3. Respondent
   a. Name and title:
   b. Contact information:

B. Substantive Questions

1. Please provide or refer to a brief description of the survey, covering its purposes, content, and design.
2. How many public-use microdata files based on this survey have been released or are expected to be released by the end of 1999? (*Note: If the answer is none, please discuss briefly why you have not released any public-use files and stop here.)*
3. For each file released, what units do the records refer to (e.g., persons, families, households, establishments)? (Some files may include more than one type of unit. List all that apply.)
4. For each file released, which of the following kinds of data does it contain: longitudinal data, linked administrative data, contextual data for small areas? *(Note: If none of the files contains any of these kinds of data, stop here. You may want to discuss briefly your reasons for not issuing any public-use files with these kinds of data.)*

Questions 5, 6, and 7 refer to longitudinal, linked administrative, and contextual data, respectively. Answer only those questions that apply to one or more of your public-use microdata files. *(Note: We understand that you may not wish to provide full detail on masking and other procedures used. If so, please indicate that some details have been excluded.)*

5. Longitudinal data

For what kinds of units are longitudinal data included in the files, e.g., persons, households, families, establishments?

What are the main categories of data and some of the key variables for which longitudinal data are included in the file, and what are the reference dates or periods for these variables?

Which of these variables were considered to present special disclosure risks and what masking or other procedures were used to reduce these risks to an acceptable level?

If age was a variable, how was this dealt with for successive time periods?

Were some variables that were included in one or more cross-sectional files excluded from longitudinal files? If yes, what kinds of variables?

Were there any special problems associated with persons or households that moved between waves of the survey or entered or left the sample for other reasons?

How were these dealt with? How much geographic detail was included in the files? Was this less than that included in cross-sectional files?

6. Linked administrative data

For what kinds of units are linked administrative data included in the files, e.g., persons, households, families, establishments?

What agencies or organizations provided the administrative data and what specific administrative record databases were used?

How were the administrative data linked to the survey data, e.g., using SSNs or other numerical identifiers, using name and address, etc.?

What match rates were achieved?

What are some of the key variables for which linked administrative data are included in the file, and what are the reference dates or periods for these variables?

Which of these variables were considered to present special disclosure risks and what masking or other procedures were used to reduce these risks to an acceptable level?

What procedures were used to prevent the supplier of the administrative data from being able to reidentify individual units in the public-use file containing the linked administrative data?

7. Contextual data (data on the characteristics of political subdivisions or other geographic areas in which the sample units are located)

What kinds of areas were contextual data included in the public-use microdata file(s)?

What kinds of variables were included for these areas and how much detail for each one? *(If possible, provide a list)*

What precautions were taken to ensure that users would not be able to use these variables to identify areas not explicitly identified in the file(s)?

Please mention any other issues concerning statistical disclosure limitation procedures used for these files that you think might be of interest for this study? For example, we would be interested in knowing how you chose the particular masking or other methods you use from among the range of alternatives available.

THANK YOU FOR YOUR COOPERATION!

Appendix 2. Availability of "at risk" data for public use files meeting study criteria.

Data set no.	Availability of longitudinal data	Availability of linked administrative data	Availability of geographic or contextual data
01	Public use	Restricted access	Public use
02	Public use	Restricted access	Public use
03	Public use	None	Public use
04	Public use	Public use	Public use
05	Public use	Public use	Public use
06	Public use	None	Public use
07	Public use	Restricted access	Public use Restricted access
08	Public use	None	Public use
09	Public use Restricted access	None	Restricted access
10	Public use	Restricted access	Public use Restricted access
11	Public use	None	Public use Restricted access
12	Public use	None	Public use Restricted access
13	None	None	Public use
14	Public use	Restricted access	Public use Restricted access
15	Public use	Public use	Public use
16	Public use	Restricted access	Public use Restricted access
17	Public use	None	Public use
18	Public use	None	Public use
19	Public use	None	Public use
20	Public use	Public use	Public use