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

# Extending Theory for User-Centered Information Services: Diagnosing and Learning from Error in Complex Statistical Data

**Alice Robbin** *
*School of Library and Information Studies, Florida State University, Tallahassee, Fl 32306-2048. E-mail: arobbin@lis.fsu.edu*

**Lee Frost-Kumpf**
*School of Public Affairs and Administration, University of Illinois at Springfield, Springfield, IL 62794-9243*

**Utilization of complex statistical data has come at great cost to individual researchers, the information community, and to the national information infrastructure. Dissatisfaction with the traditional approach to information system design and information services provision, and, by implication, the theoretical bases on which these systems and services have been developed has led librarians and information scientists to propose that information is a user construct and therefore system designs should place greater emphasis on user-centered approaches. This article extends Dervin's and Morris's theoretical framework for designing effective information services by synthesizing and integrating theory and research derived from multiple approaches in the social and behavioral sciences. These theoretical frameworks are applied to develop general design strategies and principles for information systems and services that rely on complex statistical data. The focus of this article is on factors that contribute to error in the production of high quality scientific output and on failures of communication during the process of data production and data utilization. Such insights provide useful frameworks to design a distributed system of social cognition that will detect, diagnose, communicate, and learn from error. Strategies to design systems that support communicative competence and cognitive competence emphasize the utilization of information systems in a user-centered learning environment. This includes viewing cognition as a generative process and recognizing the continuing interdependence and active involvement of experts, novices, and technological gatekeepers.**

## Introduction

During the last decade there has been an explosion of publicly available, large-scale, complex statistical data. Whether available locally on diskette or cd-rom or through local- or wide-area networks linking an information center to another site, access to extensive collections of complex statistical data has altered the information use environment.[1] The term *data complexity* is used to describe datasets derived from censuses, administrative records, and longitudinal panel surveys that may contain millions of observations, thousands of data elements with numerous interdependent relationships, and intricate logical conditions on the measurement process.[2]

Data complexity connotes an interdependence of organizational, social, cognitive, and technical requirements for understanding, retrieving, and using data. These very requirements have, however, resulted in underutilization, misuse, and nonuse of complex data by data users (secondary analysts). Utilization has come at great cost to individual researchers, the information community, and to the national information infrastructure. The knowledge requirements for handling statistical data are now so extensive that it is no longer possible for a single researcher working alone to analyze data without expending significant computational resources and personal time. There is, however, a growing recognition that using statistical data efficiently and effectively requires that the traditional relationship between a dataset and the user be altered. And now that information

* To whom all correspondence should be addressed.

Received October 16, 1995; revised December 4, 1995; accepted December 4, 1995.

© 1997 John Wiley & Sons, Inc.

[1] Following Morris (1994, p. 20, fn. 1) we use the term "information services" to cover all areas that serve users in a library or information center.

[2] What contributes to their complexity are their data structure, representation of time in measurements, analysis units and aggregations, representation in machine readable form, and the extensive amount of information required to use the data appropriately. For more discussion about the attributes of complexity, see David (1991) and David & Robbin (1990a, 1990b, 1992).

centers have become more involved in the provision of data user services, organizational structures must be designed for managing these data in order to improve their accessibility and ensure high quality scientific output. But what should these organizations look like so that complex data will be used efficiently and effectively by social scientists and others?

In recent years some librarians and information scientists have articulated increasing dissatisfaction with the traditional approach to information system design and information services provision, and, by implication, the theoretical bases on which these systems and services have been developed. They argue that professionals must move *away from* an approach that views information or data as objective, and information or data use as determined by the expert system designers and system managers as controllers of data resources. They must move *toward* an approach that posits information as a user construct and system design as a user-centered enterprise. Implied by this change in orientation is that our conceptions of users, how they perceive information, and their information needs and uses require a different theoretical basis for information system design and user services.

We strongly concur with Dervin (1977, 1983, 1992), Dervin and Nilan (1986), and Morris (1994) who believe that a strong theoretical basis must guide information center and systems design. This article extends Morris's (1994) theoretical framework for designing effective information services, which is derived principally from cognitive psychology, to embrace theory that derives from multiple approaches in the social and behavioral sciences. The discussion also extends Webber's (1991/92) and Wilson's (1995) arguments regarding the reasons why relevant information goes unused in research, development, and policy formulation efforts. Diverse theoretical foundations must be synthesized while attention is directed to multiple levels of analysis for the production of high quality scientific output. These theoretical frameworks are applied in order to develop general design strategies and principles for information systems and services that provide statistical data, in particular, those data which are designated as complex.

A useful way to develop these strategies and principles is to focus on what generates or contributes to error in the production of high quality statistical output. The *key deterrent is inadequate communication about data.* Communication failures lead to error production both by the data producing agent and by secondary analysts. Communication failures obscure the location and sources of error in the data production process and contribute to the diminished quality of scientific output. They also impede improvements in the scientific design of future data collections. Thus, as a first step towards developing organizational structures to manage complex data, it is clear that how error is produced within the processes of data production and data utilization must be understood. Developers can then apply this understanding to designing communication and information systems that help users discover, understand, and control error.

This article shows how diverse theoretical frameworks and a large body of research contribute to developing useful strategies and principles for designing a distributed system of social interaction, communication, and cognition that facilitates error detection, diagnosis, and correction for complex statistical data. Such strategies support systems in which sense-making and interpretative processes are encouraged and operate in a social context.

A distributed system of social cognition recognizes that traditional approaches to understanding data production and utilization are inadequate. The underlying assumption behind the design of nearly all organizations and information systems that produce or use complex data has been that the end user is a single, isolated individual who operates on complex data through the computer. This perspective is faulty. Data production and secondary analysis are carried out in a social context. They depend as much on social interactions and exchanges among individuals and groups as they require a range of cognitive reasoning processes. Such social interactions and exchanges require large-scale, bureaucratic organizations to organize, routinize, and manage data production activities associated with panel surveys, administrative records, and censuses. Individuals involved in data production and use must collaborate on a daily basis to complete a large variety of routine and nonroutine tasks. This means that people search for and obtain assistance from one another; they seek help and learn from peers and experts. Taylor's (1968, 1986, 1991), Taylor and Utterback's (1975), and Garvey's (1979) studies on communication behavior of scientists and engineers reinforce the contention that the proper focus must be the social situation. This includes the relationship between the user-scientist and her environment, whose major constraints and resources are other scientists and complex bureaucratic organizations that provide data, interpretations, and other forms of user services.

The dynamic and loosely coupled nature of social interactions and social cognition which underlie the data production and utilization processes create problems in understanding what the data mean and the appropriate uses to which they can be applied. This occurs because knowledge and discoveries about the data are not recorded or communicated in any systematic or shared way. Information about the data, including its uses and limitations, is often revealed only through informal dialogues and conversations.

A physical structure for this distributed system of social cognition is neither devised or prescribed. Rather, fundamental theoretical principles relevant for designing systems that more effectively and efficiently produce and facilitate using complex data will be identified and elaborated. These principles are guided, as Norman (1983a, 1988b), Senge (1990) and others have argued, by appreciation for the roles played by: 1) The mental models of those who work in organizations that produce and utilize complex data; 2) the practices of individuals engaged in data production and social science inquiry; and 3) the concepts associated with the design of information systems to collect, process, and utilize complex data that are embedded in learning organizations.

By elaborating the theoretical foundations of the social, interactive processes of data production, scientific discovery and knowledge production, information systems can be designed that will improve the quality of complex data and secondary analysis. As such, this article is an effort to use theory to design tools that can result in higher quality data and contribute to fundamental improvements in scientific practices.

Part One discusses the core assumptions, definitions, and concepts underlying a framework for a distributed system of social cognition to detect, diagnose, communicate, and learn from error. Attention is focused on different types of error that occur through various organizational, social, and cognitive processes. Understanding the different types of error and their underlying assumptions serves as a critical foundation for recommendations on design strategies and principles. In Part Two, the concepts of error and social cognition are defined, five types of error are identified, and the relationship between cognitive development and social interaction is explained. This section concludes with a summary of the assumptions and offers a series of propositions to link the assumptions about communication processes, the production of knowledge, and error. Part Three offers two general design strategies for a distributed system of social cognition to detect, diagnose, communicate, and learn from error. Our strategies to design for communicative competence and cognitive competence emphasize the utilization of information systems in a learning environment, social cognition as a generative process, and the interdependence and active involvement of experts and novices. Principles are adapted from design rules based on analyses of human error and derived from assessments of communication/information technologies to support and enhance work group performance.

Designing user-centered, distributed systems of social cognition requires theory and knowledge from several disciplines. We draw freely on and extrapolate from theory as well as field and laboratory research on organizations and social networks, scientific collaboration in cooperative work groups, learning and cognition, and human-computer interaction to examine the data production and secondary analysis of longitudinal panel surveys. The argument is located in different theoretical traditions that use different concepts and terminology to express essentially similar meanings and that emphasize different aspects of how organizations and individuals process information and construct meaning. This is not to argue, however, that the assumptions and premises of these different theories are the same. Rather, emphasized is that the understanding of information systems design will be enriched by integrating different theoretical approaches. Theory and practice are used to establish a framework for a task system of data production and utilization that maximizes the ability of people operating in that system to detect and learn from error.

Throughout this article, examples are used to illustrate theory and concepts. These examples and recommendations for developing strategies are drawn from experiences with two complex, dynamic longitudinal panel surveys that have served as sources for significant social policy decisions. The first dataset is the *Survey of Income and Program Participation* (*SIPP*), a longitudinal panel study to examine the economic well-being of Americans, conducted by the U.S. Bureau of the Census (Ryscavage, 1987). The second is the *Wisconsin Child Support Reform Program* (Garfinkel et al., 1988), a project carried out by the Institute for Research on Poverty at the University of Wisconsin-Madison, which collected and integrated data from court and other administrative sources produced by many, independent administrative units for 10 cohorts (panels) and also conducted related sample surveys for a period between 1980 and 1995. While the examples offered to support the argument are undeniably complicated, they serve as important illustrations of the statistical data that are commonly used by social scientists who are engaged in public policy research and evaluation.

## Part One. Framework Assumptions

Three basic assumptions are critical to the focus of this article and recommendations for design principles. The first assumption is that error is inevitable and almost certain to occur because data production and utilization systems are loosely coupled and because organizational, technical, and cognitive processes are very complicated. The second assumption is that data production and use are communication processes, articulated by language, which take place in a social context. Errors that occur in and about a set of data not only pertain to the errors in measurement of the data themselves, but also encompass misinterpretations and miscommunication about the data, their measurement, and their meanings. This means that people can learn about error, and that errors in the data and the data production and utilization processes can be identified, corrected, controlled, and used as a source for long-term learning opportunities.

### *Assumption 1. Error*

The complexity of the organizational, social, and technical processes embedded in data production and use means that the occurrence of error is inevitable (cf.

Norman, 1983a, 1983b, 1986a, 1986, 1988a, 1988b). Error is socially produced in the processes of data production and data utilization.

*Data production process.* The complexity of these processes means that data producers will always produce public data files that contain error. The data producer's technology and its structural properties introduce error that finds its way into public use datasets and its subsequent utilization. One example illustrates the role that organizational factors play in introducing error into the production and subsequent use of complex data.

The administrative complexity of record keeping practices creates failures in the transfer of data and information between governmental agencies. This contributes, for example, to significant errors and high rates of missing data in the *Wisconsin Child Support Reform Program* data set. Collection of child support payments and related data by the county Clerks of Courts depends on several critical steps and contingencies that required efforts of several other public agencies. Each agency has its own base of authority, power, and discretion, with its own complex set of rules, procedures, and meanings about the data and how they will be collected, processed, and used. To illustrate, wage assignment information about non-custodial parents who are legally required to pay child support is received by the county Clerk of Courts, based upon reports from employers who are required to withhold child support under Wisconsin law. Some of the decision rules and contingencies that operate to make the data complex include the extent to which: 1) The non-custodial parent is employed; 2) an employer is notified that a current or prospective employee has been ordered to pay child support; 3) an employee is paid wages or salary through an employer's payroll system or on a "cash" (non-reported) basis; 4) an employer notifies the Clerk of Courts of the wage assignment; 5) the Department of Industry and Labor Relations notifies the county Clerk of Courts about the employment status of the non-custodial parent; and (6) the county Clerk of Courts has an up-to-date record of the court order which assigns child support payments. Wage assignment information may also be received through the Wisconsin Department of Revenue or Internal Revenue Service (U.S. Department of Treasury). At several points in this process, public or private agents contribute information regarding wage assignment to child support payments. At each point, errors can occur in reporting, recording, or interpreting wage and eligibility data, including reconciling two or more "correct" versions of the data. Furthermore, considerable lag time may occur between collection, recording, and transfer of such data among administrative units. This "transfer lag" can contribute to additional error, because current information may be commingled with older, obsolete information and thus may create false, misleading, or missing information. One effect of delays in the transfer of information between administrative units is that information about the child support payment is not yet available at the time that other data collection efforts take place in the field.

*Data utilization process.* The production and use of complex data depend upon learning processes. For analysts charged with "making sense" of the data, the learning process means that analysts will always make errors. Seifert and Hutchins (1989, p. 42) contend that the fundamental reason why errors will inevitably be made is because use of the data relies on "learning on the job." And, they explain, "where there is the need for learning, there is the potential for error."

A different example illustrates how the data producer introduces the potential for analyst error. We often find that data are duplicated in public use files. For example, public use files produced by the U.S. Bureau of the Census contain variables that are used for both internal Bureau data processing and analysis and external secondary analysis by researchers. The public use files of the *Survey of Income and Program Participation* (*SIPP*) contain original questionnaire items, recoded variables of original questionnaire items, constructed variables derived from recoded or original questionnaire items, and imputation flags of the original questionnaire items and recoded variables. Many original questionnaire items also appear in edited and unedited form, but most users expect "only one variable for each concept, not two or more" (McMillen, 1990, p. 1). Complexity introduced by the Bureau's data collection and processing requirements adds immeasurable complexity to the data, its potential and often competing interpretations or meaning, and its likely uses in producing supportable, analytical results. Organizational, social, and technical processes associated with data production and utilization introduce uncertainty and create ambiguity in the meaning, interpretation, and quality of the data.

Severe limits on the abilities of human beings to process large amounts of information and to know its meanings and constraints also contribute to the production of error and to an incapacity to uncover errors. This occurs because analysts cannot remember thousands of variables and the logical conditions that apply to the measurement of each variable. For example, to create a cross-sectional sample from the *SIPP* often requires examining a series of attributes in the interview under investigation as well as in prior and subsequent interviews. Creating a cross-sectional or longitudinal analysis file also requires examining data on the interview status, sample relevance, and demographic characteristics of the respondent in every interview. Such linkages run counter to the expectation that each cross-sectional interview is independent of any prior interview. The variations in how the data were collected and the analyst's expectation of logical independence (based upon concepts of statistical independence) has consequently resulted in significant errors when analysts attempt to create extracts from these public use files.

McMillen (1990, p. 1) has also observed that analysts make errors about the different concepts of time embedded in the *Survey of Income and Program Participation.* This is because some data are collected for specific weeks, all weeks, months, two or more months, or all months in the reference period of an interview. Each time interval requires an understanding of the calendar that corresponds to the reference period in each of four independently drawn subsamples. He notes that "Disaggregating data collected for several months and aggregating data collected below the monthly data provide a variety of sources of confusion and error." Indeed, different modes of time—survey time, reference time, and calendar time—have been a continuing source of misunderstanding and, ultimately, errors produced by analysts.

Why does error occur? Norman (1983b, p. 8) concludes, after careful study of human error and human-machine interaction, that

> most people's understanding of the devices they interact with is surprisingly meager, imprecisely specified, and full of inconsistencies, gaps, and idiosyncratic quirks. The models that people bring to bear on a task are not the precise, elegant models. . . Rather, they contain only partial descriptions of operations and huge areas of uncertainties. Moreover, people often feel uncertain of their knowledge—even when it is in fact complete and correct—and their mental models include statements about the degree of certainty they feel for different aspects of their knowledge.

In a later analysis, Norman (1988b, p. 114) adds that mistakes derive from various sources, which include "poor decision making, misclassifying a situation, or failing to take all the relevant factors into account." People rely on faulty memory or remembered experiences and do not apply systematic analysis to the task at hand. Furthermore, they routinely minimize the amount of information they need to make decisions "or the completeness, precision, accuracy, or depth of the learning" that is required to perform a task (p. 55). They do, however, make use of constraints "that simplify what must be retained in memory" (p. 61). Relatedly, Simon (1979), relying on his studies into the limits of human information processing, has noted that the process of decision making is serial, selective, and satisficing. Decision makers obtain just enough information to solve the task or problem at hand. Considerable research into psychological and behavioral aspects of human judgment and decision making support this claim, as well (Nisbett & Ross, 1980; Slovic, Fischhoff, & Lichtenstein, 1977; Tversky & Kahneman, 1982).

Observations of those who work in data producing organizations and who practice social science lead to the conclusion that both producers and users generate many different errors during the various phases of data production and utilization. This occurs because of insufficient knowledge and experience needed to understand the data, faulty application of past experience to a new dataset, and the technical requirements of the tasks at hand.

The traditional response to preparing public use files illustrates this problem. The Bureau of the Census operates like most other data producers. Attributes (variables) for a given record are all organized in sequential order and represented in a standard flat file or rectangular matrix structure. This has two consequences. The structure of such files completely obscures many of the logical relationships between attributes; this often violates the intent of the original scientific design. A second consequence is that much of the data designed into a survey's complex response structure is thrown away. Much of the data that is thrown away provides information about important relationships between different sets of measurements made on different units and levels of analysis. As the Panel to Evaluate the *Survey of Income and Program Participation* noted in their final report, processing of data in the form of a flat file structure destroyed certain classes of household relationship networks created in the questionnaire which could not be recreated (National Research Council, 1993b; Olsen, David, & Sheets, 1991).

The same logic of organizing data is embedded in the *Wisconsin Child Support Reform Program,* as well. Data for the panels are organized in a rectangular matrix structure with all attributes associated with one unit of analysis, regardless of what was measured. In the child support project, this "standard" unit of analysis is the case file (court record of separation, divorce, or paternity), despite the fact that there are many different units of analysis measured as part of child support proceedings (e.g., data measured for families, children, custodial parents, non-custodial parents, and social welfare "cases"). The standard file structure eliminates essential data relationships between different measures for different units of analysis (e.g., mothers, fathers, individual children) over time. This occurs despite the fact that such data relationships already exist in the structure of court records, tax records, and social assistance case records.

Both the data producer and user experience difficulties with understanding and effectively handling complex data because they lack sufficiently well-developed conceptual schema to apply or augment existing knowledge to new situations that use complex data. Relying on existing rules developed through past practices with other types of data often presents constraints that do not help them understand the structure and meaning of new forms of data. Furthermore, the production and use of new types of data require that both the data producer and data user solve complicated (e.g., ill-structured and unstructured) problems.[3] Few data producers are, however, experienced with producing longitudinal surveys or preparing data from administrative records. In general, both producer and user assume that the same set of rules and procedures developed from past practices apply to new situations involving new types of data. In many cases, this means that they incorrectly apply past rules and procedures to process longitudinal data that are based on their experiences with cross-sectional survey data. Similarly, researchers make errors because their tools, experiences, and specific knowledge pertaining to a data set are inadequate.[4]

*Assumption 2. Communication*

Language organizes the categories of our social reality and structures the way in which we approach situations. Language communicates our interpretations of the empirical world and allows us to express our schema or mental model about objects, events, and people that evoke our attention (Rogoff, 1990; Schutz, 1962, 1964, 1967). Scientific concepts and technical terminology are used to communicate about complex data. This form of communication includes verbal and nonverbal cues that help organize useful categories for creating meaning and understanding based on individual and collective constructions of social reality.

*Interpretive processes.* Data are an important product of an interpretive process in which language forms are converted from expressive utterances or discourse to symbolic representations. The conversion from discourse to symbolic representations (such as symbols and text) involves socially active individuals interacting through various conversational modes to establish a consensus on the meanings, significations, and signs attached to the production and use of data.

Interpretation is "inextricably interwoven with the context" of data production, which "includes [the] physical and conceptual structure of [members of the organization] as well as the purpose of the activity and the milieu in which it is embedded" (Rogoff & Lave, 1984, p. 2). For example, rules and procedures for coding data from administrative records by a data collector or for judging the acceptability of a response from a human subject by an interviewer will reflect certain pre-established and usually learned categories for organizing the collection of the data. These rules, procedures, and categories are designed to eliminate ambiguity in the interpretative process.

Two examples illustrate these situations. For more than 5 years, coders for the *Wisconsin Child Support Demonstration Program* systematically ignored data that did not fit the pre-established categories created by the research team for the design of the initial data collection. It was only when the amount of data not being coded increased significantly, that some coders began to question the stability of the rules and procedures and the effectiveness of the pre-established categories. Substantial ambiguity had thus been introduced into the data collection process. As with other data produced by the U.S. Bureau of the Census, researchers who use the *Survey of Income and Program Participation* possess virtually no information on "Don't Know" responses. This is because the Bureau has a policy of not allowing "missing data" in public use files. (This is accomplished by estimating valid values for missing data based upon the data values for similar, matched, sets of records during the data processing stage.) The Bureau's editing system also eliminates response inconsistencies (e.g., eliminating variations in a subject's response to questions about gender, race, age, or marital status during subsequent interviews). This occurs even if an inconsistency makes sense within a particular context.

These two examples also illustrate the length of the communication chain for the data production process, which involves data collection, entry, and conversion of response data to machine readable form, editing and revision, and storage, retrieval, and management (Clark, 1986). At each stage in the process, interpretation, sense-making, and construction of meaning are required to generate a shared sense of understandings about what the data represent and how they can be used. Interpretation, sense-making and construction of meaning will vary, however. They will be different because the production of complex data necessarily requires contributions by different agencies and staff with different duties, tasks, authority, and responsibilities. Understandings and knowledge about the data, what they mean, and how they can be or are used, will differ according to the char-

[3] We rely on Mintzberg, Raisinghani, & Theoret, (1976, p. 246) for the definition of unstructured. Unstructured refers to "decision processes that have not been encountered in quite the same form and for which no predetermined and explicit set of ordered responses exists in the organization." Winograd and Flores (1987, p. 153) explain that structured tasks are those that follow a set of rules that can be programmed by a computer, but no such rules exist for unstructured problems. They use the term semi-structured to define those tasks which "recur but not so much that one can fully specify the relevant rules."

[4] Both the data producer's and researcher's capability to avoid error requires three different domains of knowledge: Tools (programming languages, statistical programs, database languages, computer operating systems); experience (survey design and field problems, administrative record systems, very large databases, published empirical studies); and, specific knowledge (information pertaining to the data set acquired from past work on the data or prior work as the data producer (see David & Robbin, 1992, vol. 2, pp. 21-24). In an earlier article by David (1980, p. 329), he adds that a knowledge of computer science is "helpful for clarifying the logical structure of the data and implications of that structure for analysis," in addition to improving upon efficiency through particular types of programming, and "techniques for assessing the numerical accuracy of our results, so we can avoid the delusion of accuracy in a set of results." Since we do not expect that this fourth domain of knowledge is part of the knowledge base of the social scientist, our design strategies discussed below embody this knowledge in an expert.

acteristics of the tasks performed by each agency or staff member.

Thus, for example, administering a large-scale national survey like the *Survey of Income and Program Participation* is a complicated organizational process, involving many different organizational units and staff distributed over many sites. Many different units at the U.S. Bureau of the Census are necessary to manage the production of public use files. One unit is responsible for the scientific design of the survey. Another unit is responsible for data collection. Other units are responsible for assessing data quality and for entering and processing (transforming) responses obtained in the field into machine readable form. Another unit develops the algorithms that recode variables, impute missing items, statistically match missing interviews, and create aggregations for new types of analysis. And yet another unit, either inside the Bureau or as an independent organization, prepares and distributes public use files and their documentation. Thus, a great deal of information does not get communicated clearly, consistently, or in a timely manner across these organizational units.

Furthermore, members of these units may have little or nothing to do with each other because the tasks related to data production are carried out separately and semi-autonomously in a loosely coupled organization with little or no feed-forward or feedback processes. A direct consequence of this loosely coupled, semi-autonomous data production system is that it substantially contributes to the production of error, including not documenting the data and the data production process. It is important to emphasize, however, that the production of error is not simply a problem of poor performance or lack of training or experience on the part of staff. Rather, it is a natural part of the social context and social milieu, reflecting complex, yet loosely coupled interdependent relationships over many sites.

Interpretations, understandings, and knowledge of the data generated by each organizational unit are not institutionalized as part of the data producer's repertoire of competencies. These understandings do not become part of the data producer's "memory," and are therefore permanently lost (Dolby, Clark, & Rogers, 1986). For example, many processing decisions related to the *Survey of Income and Program Participation* were never formally recorded by members of the U.S. Bureau of the Census outside the code embedded in editing programs. Even analysts inside one administrative unit could not obtain information about decisions made by programmers who were located in another unit. Decisions about which data were excluded from the public use files but which had appeared in the original survey instruments were not formally documented for secondary analysis. Descriptions about recoded variables and about most imputations of data values to missing data or missing interviews were never made (and, we suspect, never documented). Analysts were never informed that many variables were unedited. Releases of corrected public use files were not accompanied by information on which observations, variables, or values were modified.

Yet another, different, interpretive process occurs when the data are used. Usually far removed from the organizational routines of data production, users make inferences in order to make sense of the data. Like members of the data producing organization, their background knowledge is a product of prior cognitive skills, motivation, social experiences, and the problem-solving context (Butterworth & Light, 1982; Hannaway, 1989; Hastorf & Isen, 1982; McPhee & Tompkins, 1985; Mitroff & Mason, 1981; Roloff & Berger, 1982). Interpretations will therefore differ between users and members of the data producing organization, because differences in accumulated knowledge derived from everyday experiences rely on a different language about data and reflect different goals and problems.

The implication is that a "common ground" (Clark & Carlson, 1982) of understanding cannot be assumed and may not be possible. Members of data producing organizations and users have different understandings of the data that are grounded in different assumptions—what Krauss and Fussell (1990) call the "mutual knowledge problem."

Because communication failure characterizes both the data production and use environments, it results in reduced effectiveness and inefficiency in the scientific process. Trial and error prevail as users make discoveries about the quality of the data. Yet, as this source of expertise develops, users who have investigated and learned about the data do not have a vehicle for collating and communicating knowledge and expertise derived from their experiences. New cohorts of users encounter the same problems and make the same discoveries repeatedly. Because there is little or no feedback communication with the data producer, documentation and improvement to scientific designs and data collection procedures do not occur.

A distributed task system of social cognition therefore depends on the ability to "construct a common cognitive environment" inside the data producing organization and between the data producer and user, that is, to "ascertain and represent the information that. . . can (and will) be assumed to be known to all" (Krauss & Fussell, 1990, p. 112). As Rogoff and Gardner (1984, p. 97) comment, "People who are concerned with jointly accomplishing a cognitive performance must possess or create a common framework for the coordination of information."

*Storage and retrieval of communication as information: Cognitive structures.* The complexity of the data comes about, in part, because the elements of a dataset are closely related to one another. That is, decisions are made about one element without taking account of other, related elements, and thereby will lead to the occurrence of error. For example, it is easy to make a significant error in extracting a subsample of the population in the *Survey of Income and Program Participation.* The analyst cannot extract individuals based only on the (i) interview status (e.g., adult, child, proxy, etc.) for a particular interview without knowing whether the individual was (ii) interviewed during the particular time period under investigation, and (iii) sample relevant for that time period (i.e., months during which an individual was properly part of the sample). Retrieving only on (i) will almost always result in an incorrect subset of cases that will include some individuals who were not actually interviewed (e.g., the cases were imputed or no information was ever collected) and will also exclude other individuals who should have been interviewed for that temporal period.

Because the information requirements associated with using complex data are extensive, a significant amount of information must be investigated for decisions related to extracting cases and measures and conducting analysis. Conceptual complexity requires a significant search through multiple paths, as well as trial and error. A critical element of the task system is, therefore, organizing information to inform the researcher about how to avoid error.

How do we organize the structure of information about error to make it accessible, reduce uncertainty and ambiguity, focus attention, provide relevant cues, and avoid information overload? For some insights into how to organize the structure of communication/information about error, theory and empirical research in cognitive psychology draws attention to different aspects of information acquisition, processing and use, and the limitations in human information processing [Norman, 1988b; Simon, 1979; and others (cf., Hippler, Schwarz, & Sudman, 1987)]. Cognitive psychology reflects a clear commitment to the information processing paradigm, which, as Bodenhausen and Wyer (1987, p. 7) note, has not been without critics for its theoretical utility in "predicting phenomena of the 'real world' outside the laboratory." This paradigm should not be rejected, however, because it has potential utility for understanding how information is processed and how it may be more effectively organized and interpreted.

This theory makes the following argument about human information processing and organization: The two organizing concepts for memory structures that store and retrieve information are short-term memory (STM) and long-term memory (LTM) (Simon, 1979). STM can only hold a limited number of chunks of information, but the number of chunks can be increased by practice (Goleman, 1994). Simon (1979, pp. 41-42) explains that LTM may be thought of as an encyclopedia, with a text and an index; information is accessed via pointers that comprise the index. The text has an associative structure—"a system of nodes interconnected by numerous links. Information can be retrieved from it not only via the index but also by following paths of links from one node to another through intermediate nodes." Recognition is retrieval using the index, whereas association is retrieval using sequences of links. Association is slow, but the power to discriminate and recognize can be increased by elaborating the index. Information for assisting in the decision process can be organized so that it has a wide or narrow and deep or shallow tree structure (Norman, 1988b, pp. 121-123). Most everyday tasks are shallow and narrow: They require little planning, rely on trial and error, and have few decision alternatives. In contrast, decisions about complex data can be said to have a wide and deep decision tree structure. They require prolonged search, evaluation, extensive planning, trial and error, and problem solving (careful reasoning and thinking through alternatives). The tasks are difficult. Decisions are not straightforward and "answers are not readily deducible" (Norman, 1988b, p. 125).

What are the implications for organizing information about complex data? Shallow and narrow structures will minimize the mental computations required for searching, retrieving, and evaluating information. As Norman (1988b, p. 125) suggests, "If a structure is shallow, then width is not important. If the structure is narrow, depth is not important." The model of bounded-rationality can be applied as a guide to organizing information: Follow an incremental strategy of "dividing the difficulties at the outset and attacking them piecemeal; this is a cumulative strategy, parsimonious in its use of mechanisms and inhospitable to ad hoc solutions" (Simon, 1979, pp. x-xi). And, finally, tools can be developed that assist in making associations and indexing information to improve the retrieval process.

*Communication network structures.* What types of communication network structures can be created for effective and efficient interpersonal and group communication? Four different research programs on communication yield insights into network structures, although they make different assumptions about information, users, and the information use environment (see Lievrouw, 1988, for a discussion of these assumptions). These include empirical evidence from the late 1940s to the early 1960s, based on the work of Bavalas (1950),[5] Guetzkow and Simon (1955), and other experimental evidence summarized by Shaw (1978); studies of communication in the scientific community by Paisley (1965, 1972, 1984), Garvey and colleagues (1972a, 1972b, 1972c, 1972d), Garvey and Griffith (1974), Hagstrom (1965, 1970), Brittain (1970), and Eradi and Utterback (1984); studies of the role of information or technological gatekeepers (Allen, 1970, Katz & Tushman, 1979; Tushman, 1977; Tushman & Katz, 1980);

[5] Griffith and Miller (1970) and Hesse et al. (1993) also found that productivity was associated with extensive communication networks and a gatekeeper role.

and naturalistic observation and survey research on innovation diffusion and adoption (see, for example, Finholt & Sproull, 1990; Finholt, Sproull, & Kiesler, 1990; Rogers, 1994; Rogers & Kincaid, 1981).

Communication patterns within groups influence user satisfaction, effectiveness, and efficiency with regard to the distribution of information and task completion. The early laboratory experiments demonstrated that the network's structural characteristics, such as centralization and decentralization, had major consequences for communication and task performance. Centralized networks appeared to be more efficient when they only required the collection of information in one place, whereas decentralized networks were more efficient when tasks were complex and required multiple operations on the information before the task is completed. The hub and spoke form of communication appeared from the experimental evidence to provide an efficient structure for communicating, collating, and distributing information. The efficiency of this structure may be explained by the key role played by a "technological gatekeeper" or boundary-spanner on whom project groups rely heavily for information and who contributes to an organization's effectiveness by filtering and channeling external technology and information into the organization (Katz & Tushman, 1979). The boundary-spanner serves as a mediator between "organizational colleagues and the world outside and effectively couples the organization to scientific and technological activity in the world at large"[5] (Allen, 1970, p. 192). It appears that gatekeepers do not suffer a "communication impedance" that makes communication across information boundaries "relatively difficult and prone to bias and distortion" (Katz & Tushman, 1979, p. 143).

Two criticisms were subsequently lodged against this early laboratory work on communication network structures. First, it assumed a Shannon-Weaver linear model of communication, which was vigorously criticized for its underlying assumption of one-way causality of the components of the model on communication effects. Rogers and Kinkaid (1981, pp. 34, 37), for example, argued that this model described the "act" not the process and was, furthermore, "atomistic and mechanistic." (See also Finholt et al., 1990 for a critique of this early model of communication.) Instead, the extensive research on innovation diffusion and adoption, by fusing systems theory and social network theory, provided an alternative model of communication that emphasizes the "mutual sharing of information in order to achieve some common purpose, like mutual understanding and/or collective action" (Rogers & Kinkaid, 1981, p. 31; see also Rogers, 1994).

Second, as Finholt and colleagues (1990, p. 297) observed, the electronic network creates a different environment than the laboratory setting because electronic networks provide flexibility and can support different forms of communication.[6] Furthermore, they hypothesized that the networks might offer a "new opportunity to participate in larger, less structured and heterogenous groups than the work group" and might give people a feeling "of increased participation and therefore increased commitment to the organization as a whole" (Finholt & Sproull, 1990, p. 60). Indeed, evidence from other projects appears to support their hypothesis that project success is related to communication and that the creation of new knowledge derives from the rapid sharing of expertise across subunits and across work groups regardless of geographic proximity (see Comer, 1983; Cotter, 1988; David & Robbin, 1992; Hesse, Sproull, Kiesler, & Walsh, 1993; National Academy of Sciences et al., 1989; National Research Council, 1993a; Robbin, 1992, 1995; Schatz, 1992; Steele, 1984). Thus, geographic proximity may be less important than earlier postulated (but for an opposing argument about geographic proximity, see Kraut, Egido, & Galegher, 1990). This finding was also supported by research carried out by Eradi and Utterback (1984), who examined the effects of communication on technological innovation and found that the frequency of communication in scientific projects was positively correlated with project success.

These research findings on communication structures and boundary-spanning individuals have relevance for the design of a distributed task system to reduce error resulting from the production and use of complex data. These findings suggest that a task system must meet the functional requirements of coordinating communication among its members, providing channels for communicating, and providing access to and distributing information. How these functions are performed will vary across time, contexts and individuals, depending on the nature of the task, group, and organization/institution within which the individual and group are located.[7] These findings also suggest that we can increase efficiency in communications about error by centralizing information about error. We would then expect a reduction in uncertainty, ambiguity and equivocality, as a result. And we can increase the effectiveness of the research process by institutionalizing an expert gatekeeper who

[6] They note that, "Conditions are quite different in a computer communication system being used to support an ongoing group task. In particular, although all users of such a system are completely interconnected, the flexibility of these connections allow groups, at different times to create structures that meet their functional needs. This is an important distinction between these systems and the crude lab apparatus employed in the earlier studies. It suggests, further, that there are limits to the application of the structural metaphors from these experiments. That is, a system which at one point assumes a functional pattern resembling a spoke and hub, and at another point assumes a functional pattern resembling the circle form cannot really be said to possess the characteristic shape of either of these patterns. Instead, modern communication systems seem to produce amalgamated forms."

[7] See McGrath (1984, 1990) for discussions on groups and how time affects the content and flow of information.

collates, evaluates, synthesizes, and communicates information about the data to members of a decentralized communication network.

### *Assumption 3. Knowledge*

Our third assumption is that knowledge is socially produced and derived (see Hutchins, 1991; Lave, 1988; Rogoff, 1990; Rogoff & Lave, 1984). Knowledge about data and how to avoid making errors is distributed and distributed differentially across persons, activity, and setting. Hutchins (1991, p. 306) observes that, "The tasks of learning, remembering, and transmitting cultural knowledge are inevitably distributed. The performance of cognitive tasks that exceed individual abilities is always shaped by a social organization of distributed cognition." He points out that, "All division of labor requires some distributed cognition in order to coordinate the activities of the participants" (p. 284).[8] How a group operates will depend on the distribution of knowledge among its members.[9]

Decisions about producing data—which represent embedded knowledge within the institutional context—are socially distributed across participants and through time (Mehan, 1984, p. 63). Decisions, Mehan notes, are arrived at through many small actions, "distributed across many different levels of an organization [and] are the consequence of routine, standardized procedures" (p. 65). March (1991, p. 73) echoes Mehan, "Organizations store knowledge in their procedures, norms, rules, and forms. They accumulate such knowledge over time, learning from their members." This knowledge is stored in the heads of the organization's members, whose career trajectories may change and who may depart. The organizational memory is therefore vulnerable. Consequently, knowledge must somehow be retained, permanently stored, and capable of being retrieved, in order for it to be transmitted to new members (Simon, 1991).

Another aspect of the social distribution of knowledge is the relationship between the novice and expert. Knowledge is acquired by the novice from an expert, inside the data producing organization, in the classroom or in work situations, informally and formally, through tacit and explicit forms of observation, instruction, guided participation, and experimentation. There is, however, significant variability in the way novices and experts represent, search, and retrieve information, and in their modes of and success in problem solving (Bateson, Alexander, & Murphy, 1987; Carlson, 1990; Chase & Simon, 1973; Larkin, McDermott, Simon, & Simon, 1990; Payne, 1980; Reif, 1980; Saracevic & Kantor 1988a, 1988b; Simon & Simon, 1978). Nevertheless, the relationship between novice and expert requires a partnership of "common language and system of ideas" and the "granting of reciprocity [and]. . . a consideration of alternative perspectives" (Rogoff, 1990, pp. 148–49).

The metaphor of scaffolding is used to describe how the individual acquires skills during the learning process. Greenfield (1984, p. 118) explains that this metaphor is a theoretical model of the ideal role of the teacher (expert), who "intervenes selectively, structuring an interaction by building on what he or she knows the learner can do." Expert guidance provides the institutional or cultural memory, transmitting the relevant contextual knowledge and cognitive skills required for the task (Hutchins, 1990; Rogoff, 1990; Rogoff & Gardner, 1984; Seifert & Hutchins, 1989, 1992). Errors made by the individual become a signal to the expert to modify the support structure of the scaffold (Greenfield, 1984, p. 136).

Seifert and Hutchins (1989, 1992) point out that there is a career trajectory related to individuals who are part of a collaborative work group. Less skilled individuals gain access to and acquire information, becoming more skilled in the process. Based on their study of how novice sailors in the U.S. Navy learn how to navigate, they suggest, "As a consequence of this alignment of career trajectory with the path of information through the [task] system, if one has access to an error, one also has knowledge of the processes that may have generated it, because one has already—at an earlier career stage—performed all those operations" (p. 44).

Peer interaction also contributes to cognitive development. Rogoff (1990, p. ix) notes that while peers may be less skilled, they "may offer unique possibilities for discussion and collaboration." Their unique contributions include "motivation, imagination, and opportunities for creative elaboration of the activities of their community." Rogoff also points out that those peers who are relatively more skilled find effective ways to "achieve shared thinking that stretches the less skilled partner's understanding" (p. 39).

But no matter whether the member of a data producing organization or researcher is a novice or expert, avoiding errors made with complex data requires the acquisition of a significant amount of new information which must be organized for successful problem solving. New information will be searched and retrieved and understandings will be constructed from the perspective of the individual's existing knowledge store, beliefs, experi-

[8] Hutchins classifies the types of cognition needed: To perform the task (individual level) and to coordinate the activities of the task (group level).

[9] Hutchins (1991, pp. 284–285) cites Roberts (1964), an anthropologist whose investigations of four American Indian tribes led him to speculate about differences in "retrieval efficiency from cultural memory." "Roberts attributed differences in retrieval efficiency at the group level to variables such as group size, the pattern of interactions among individuals, the distribution of knowledge, and the time course of interaction" (p. 285). Hutchins suggests that "differences in the cognitive accomplishments of any two groups might depend entirely on differences in the social organization of distributed cognition and not at all on differences in the cognitive properties of individuals in the two groups" (p. 285).

ences, and self-perceptions using a set of heuristic rules (Linn & Clancy, 1990; Newell & Simon, 1972; Tversky & Kahneman, 1974, 1982).[10] These heuristic rules include saliency (used to select information), availability (to recall information), representativeness (to classify information), and anchoring (to retrieve initial judgments) (Mehan, 1984, p. 61, citing Tversky & Kahneman, 1974; Nisbett & Ross, 1980). Knowledge will typically be acquired through a trial-and-error approach.

Cognitive psychologists who study "mental models" or "mental maps" and problem solving by physics and engineering students offer us insights into how information and experiences are organized for problem solving and knowledge creation.[11] Mental models provide a "schema" that simplifies learning and helps us interpret a new situation (Linn & Clancy, 1990; Senge, 1990). As Norman (1988b, p. 71) explains, "The power of mental models is that they let you figure out what would happen in novel situations. Or, if you are actually doing the task and there is a problem, they let you figure out what is happening." The memory task is tremendously simplified if there is a "sensible structure. . . that corresponds to knowledge that we already have, so that new material can be understood, interpreted, and integrated with previously acquired material" (p. 69).

Applied to this discussion about complex data, members of organizations and researchers have knowledge about data and how to avoid making errors. This knowledge, acquired through a discovery/learning process that is collaborative in nature, can be incorporated in a distributed task system designed to reduce error. Although Seifert and Hutchins (1989, p. 45) contend that the distribution of knowledge means that most errors will be caught ensuring this for complex data means that knowledge about error must be institutionalized and archived. It must be recoverable and be able to be transmitted. Four aspects of knowledge/learning must be accounted for: Knowledge acquisition, information distribution, information interpretation, and organizational memory (Huber, 1991). The task system must also incorporate opportunities for data producer and peer collaboration in shared discovery. There must be an expert who provides the "scaffold" and collates, archives, and communicates their discoveries and the accumulated knowledge of organizational units, groups, and individuals involved in data production and use.

## Part Two. Concepts and Definitions

This section provides definitions for two core concepts, error and social cognition. Five types of error are identified: Inferential error, residual or model error, decision making error, methods error, and task performance error. Social cognition means the relationship between cognitive development and social interaction.

### *Error*

The definition and everyday use of the term error is to a large extent contextually dependent. The concept of error means different things to different people in different situations. Most scientists define the concept of error in fairly rigorous terms, and much attention has been devoted to developing procedures to understand and minimize the occurrence of statistical error (e.g., Cohen, 1988). The ways in which the concept of error is used have not, however, been the focus of much attention by the scientific community. Few resources have been devoted to identifying the types and sources of other errors that may impact on statistical error. Groves (1989) acknowledges that observational and nonobservational errors do not represent all sources of error in survey data.[12] He writes,

> The most notable omissions are those arising after the answers to the survey questions have been obtained by the interviewers—the coding, editing, imputation, and other data processing activities that follow the data collection phase. . . These result from actions of processors and analysts of the data (p. 12).

[10] We make note of the important role that motivation plays in information acquisition and receptivity to applying new and different strategies for problem solving. We assume for the purposes of this article that researchers are highly motivated to use complex data, but there is great variability in receptivity to adopting different strategies for problem solving. See also Spitzberg's (1987) discussion on the requirements for communication competence (see below, Part Two, on "communicative competence").

[11] Mental models are the "mental representations that underlie everyday reasoning about the world" (Garnham, 1987, p. 152). Garnham notes, however, that there are two uses of "mental models." It may mean (1) what people really have in their heads and what guides their use of things" (Norman, 1983b, p. 12), or (2) the form of the content. Garnham explains that for cognitive psychologists like Norman and others (e.g., Gentner & Gentner, 1983, and Larkin et al., 1980), mental models are "structures in semantic memory that are used to interpret a wide variety of events, part of the background knowledge we use to understand what is happening around us" (p. 152). These researchers study "what is represented in mental models." Norman (1988b, p. 38) elaborates: "Our conceptual models of the way objects work, events take place, or people behave, result from our tendency to form explanations of things. These models are essential in helping us understand our experiences, predict the outcomes of our actions, and handle unexpected occurrences. We base our models on whatever knowledge we have, real or imaginary, naive or sophisticated." In contrast, for Garnham who studies language understanding, mental models are "structures created during the comprehension of particular texts. . . [and] held in episodic memory." Garnham's interest is "more on the form [of representation] and the way [mental models] are constructed and manipulated, than. . . on their content." Our use of mental models conforms to the "what" rather than the "form."

[12] Groves (1989) classifies the types of error in survey design as observational and nonobservational error. Observational error includes interviewer, instrument, respondent, and mode of data collection. Nonobservational error includes coverage, nonresponse, and sampling.

These other types of error are more pervasive, less understood, and more elusive to define than the statistically-based concept of error. More importantly, however, experience suggests that these less well understood errors impact on the extent and type of statistical error that remains hidden within the residual error term of models developed by users working with secondary data sources. This is unfortunate, because the preoccupation of the scientific community with the statistically-based concept of error draws attention and resources away from these other possible sources of error that occur in the processing of data for subsequent use in secondary analysis.

There are at least three other types of error which are related to the production and utilization of data sources. We begin the discussion by quickly reviewing the statistically-based concept of error (inferential, residual, or model). This is then followed by a discussion of three other types of error and their likely impacts upon statistical error: Task performance error, decision error, and methods error. Examples drawn from the *Survey of Income and Program Participation* (*SIPP*) and the *Wisconsin Child Support Reform Program* are used to illustrate each type of error, how and when such errors occur, and the way in which each error impacts upon the statistically based concept of error.

*1. Inferential error.* In "classical" statistics, error is an incorrect judgment or inference based upon an obtained statistical result conditioned on the assumption that the null hypothesis is true. An error occurs when the obtained statistical result produces one of two, mutually exclusive, but incorrect decisions. In the first case, the obtained statistical result leads to the incorrect rejection of the null hypothesis when, "in reality," the conditions prescribed by the null hypothesis are true and we have simply obtained a rare or extreme result. Put another way, this is the decision that an effect does exist when, in reality, it does not. This situation involves the familiar Type I error in classical hypothesis testing designated by alpha ($\alpha$).

A Type I error could occur with the Wisconsin Child Support Reform Program data when attempting to estimate the confidence interval for the average annual net child support amount received by custodial parents. The chance of a Type I error can increase if sampling procedures do not take into account important differences in income sources and eligibility status for custodial parents receiving child welfare benefits and those not eligible to receive such supplements. An adequate sample design must take into account that someone's eligibility status may change over time. For example, many custodial parents who are eligible to receive child welfare benefits at the time child support is awarded by the court may become ineligible to receive welfare benefits when child support payments are received in subsequent months. Given the administrative rules associated with child welfare benefits in most U.S. states, the actual net amount of benefits received by the custodial parent family may be less than the amount reported on court records at the time the child support award is established. The converse also holds true for custodial parents not eligible to receive welfare benefits at the time of the court ordered child support award. If child support payments are not received in subsequent months, previously ineligible custodial parents may become eligible for child welfare benefits. Both cases would significantly affect estimates of the average annual net child support amount received within and between these two groups. Much of the problem can be traced to the unstable and dynamic nature of income flow among poor and working class households usually headed by one female custodial parent. Using administrative records that represent income and child support for a single point in time falsely portrays income as a stable quantity when it is not. Therefore, the standard errors for average income comparisons between custodial parents who are eligible for child welfare benefits versus those who are not, would be much smaller than would otherwise be expected.

The second type of inferential error occurs when the statistical result obtained leads to the incorrect decision to "not reject" the null hypothesis when, "in reality," the conditions prescribed by the null hypothesis are not true. More simply stated, this is the decision that "no effect exists" when, in reality, "the effect" does exist. This is the less-often-considered Type II error.

An example of a Type II error might include calculating the distribution of differences between the "total amount of pay earned last month" in the SIPP and earnings reported 12 months earlier. David (1991, pp. 96–97) explains that the analyst is unlikely to calculate the correct value because the procedures for collecting and processing these data are not apparent. Some procedures alter the data, such as imputation of missing data; some procedures affect the sample of information obtained, such as weighting of the data; and other procedures create unresolvable ambiguities, such as use of bi-weekly pay periods to estimate income. The general effect of these procedures is that the mean is biased toward the sub-sample of valid (non-missing) responses. Biases are introduced both by imputing data for item nonresponses and for entire records where data are missing (using a procedure called statistical matching). The bias is more pronounced for specific subsamples, such as low-income people, who experience greater attrition in the panel. Furthermore, the effect of statistical matching on the sample of low-income persons (or households) is to bias monthly income upward because the data values are obtained from individuals who have a similar demographic profile, but represent a different (i.e., more stable) earnings structure than those households for which income data are missing. Therefore, annualized comparisons of changes to income among low-income households, in particular, would tend to show a higher degree of stability than actually exists in reality. The mean tends to be upwardly biased and does not reflect true changes or instability due to statistical matching. The variance would tend to be small due to a small mean and to limits on the coding of income to avoid violating records confidentiality of specific "high income" individuals. Therefore, changes in annual income of low-income households would tend to reflect an upwardly biased mean with small variance and little change. In effect, the null hypothesis of no change in income would not be rejected when in reality substantial instability and change in income may exist (i.e., a Type II error).

*2. Residual or model error.* Statistics provides another concept of error related to multivariate model building. Here, error is viewed as the residual variation in the expected versus observed "scores" on one or more dependent or response variables under a specific statistical model that incorporates one or more independent or explanatory variables. This concept of error "as a residual" highlights possibly important sources of variation which the independent or explanatory variables in a statistical model cannot explain.

For example, there are many different variables that reflect different sources of variation, which may or may not be incorporated into a statistical model. Some of these include unmeasured changes in social program benefit eligibility rules and criteria, household composition affecting program eligibility or benefit levels, or statewide administrative decisions regarding eligibility of specific groups of households. In addition, the baseline criteria for determining eligibility can vary across social programs and often does not correspond to an intact household. Differences in everyday interpretations made by hundreds of agency staff regarding what particular data fields represent and under what conditions the data are or are not recorded increase the probability that a substantial amount of "noise" exists in the data. The implications for producing large but irreducible unexplained variation under any statistical model are significant.

*3. Task performance error.* In the field of work measurement an error is sometimes defined, if only implicitly, as a mistake made while performing some task which produces some unwanted or undesirable outcome or consequence (cf., Siefert & Hutchins, 1989, 1992). Applied to the design and management of survey research, this conceptualization of error is closely related to Groves' (1989) view that errors can be expected to occur during the processing of survey data or administrative records.

For example, the *Wisconsin Child Support Reform Program* obtained income data from the Wisconsin Department of Revenue (DOR) data files of annual income tax returns. These data were used both to supply missing income data in the court records that constitute the core of the data collection for information on divorces, separations, and paternity cases, and to examine the effects of one of these interventions in the experiment. Beginning in 1986, taxpayers could file as either single, married and separate, or married and joint. DOR organized the data file so that only one record was generated for a couple, and that record was filed under the name and social security number (SSN) of the first person appearing on the return.[13] The project staff supplied a list of SSN for whom a SSN was available through the court records. The program written by DOR conducted searches that keyed off the primary taxpayer's SSN field only, and not that of the taxpayer's spouse. One consequence of the programmer's error was that complete individual and household income for instances of remarriage in divorce cases and marriage in paternity cases could not be obtained.

Similarly, court records on marriage, separation, and divorce supply vital information about the employment history of the parents. The data collection instrument records the employer and dates of employment of the noncustodial parent as the different events occur over time (the time period between a motion to file for separation and a divorce decree can be less than a year or considerably longer). During this period, employment status and location may change many times. However, the data collection instrument records only one instance of employment status, without dating the time period for that particular instance. Consequently, an analyst is unable to determine whether nonpayment of child support is due to a change in employment status or for some other reason (e.g., administrative failure or delays in recording payment, due either to change in employer or transfer of payment information to the county Clerk of Courts record keeping system).

*4. Decision making error.* In individual and group decision making, an error is often viewed as an incorrect selection of an alternative course of action from a set of two or more alternatives. The error occurs when the selected course of action proves to be vastly inferior to other alternatives that were known or could have been made readily available at the time of the decision (Janis & Mann, 1977; March & Simon, 1958; Wilensky, 1967). Analogously, in survey research, this type of error might be best represented by a poor choice of a research or sampling design.

An example of a decision making error is illustrated with the choice of a database management system (DBMS) for processing longitudinal or cross-sectional data files. This is a choice that many data producing organizations must face. For example, the Bureau of the Census made the decision to utilize data processing sys-

[13] The description of this example is drawn directly from McCall (1989), "DOR Data Documentation." See also Phillips & Garfinkel (1992).

tems designed for handling cross-sectional (one point in time) data to a whole new set of data production activities. These new data production activities involved longitudinal (multiple points in time) data for the SIPP project. This decision had widespread effects on data quality and timeliness of public use file releases (National Research Council, 1993b). One consequence was that longitudinal data files could not be prepared until the entire panel was complete. Another consequence was that there was no easy way to edit data retrospectively or prospectively, even though the design of the survey required knowledge of whether the sample person was in the panel after a current interview. Olsen et al. (1991, pp. 2-3) also note another aspect of data processing decisions that had implications for SIPP data quality. Missing data in an interview were imputed from records of a similar respondent, but not from data for the missing respondent from his adjacent interviews. They comment that,

> the current system can generate substantial instability for a respondent's history simply because missing data from the middle of a series is supplied from a different person. Users are often unaware of this problem since the imputed values are incorporated into the raw data and signalled with an imputation flag . . . Because a primary focus of the SIPP is transitions, the danger of using imputed variables in event histories [i.e., longitudinal data on the occurrences of events] are especially great.

*5. Methods error.* Another concept of error involves the mismatch between analytical methods available to or used by organizational decision makers and their identification of the fundamental underlying structure of organizational problems (ranging from ill-structured to well-structured). Under this view, different analytical methods may be applied to generate problem statements or to produce solution alternatives that match the "true" defining attributes (or structure) of the problem at hand. An error occurs when analytical methods are used that are not appropriately matched to the problem at hand. In other words, the method selected is not "analytically congruent" with the defining attributes of the problem (Sutherland, 1978).

For example, the Census Bureau's choice of traditional computer programming methods to produce data from the SIPP longitudinal survey had repercussions for the scientific design and usefulness of the SIPP survey data. This survey was conceived as a tool for measuring economic well-being and social program participation, with a focus on short-term events. The instrument was designed in part to respond quickly to new policy issues on the political agenda. Thus, although the instrument could be redesigned to incorporate new questions, the decision to invest heavily in traditional programming solutions meant that the data processing system could not meet the research design requirements for greater flexibility. In effect, the data processing system altered one of the original goals of the scientific design of the survey in a way that made the survey's data less useful.

### *Social Cognition*

The focus of attention is on the relationship between cognitive development and social interaction (cf., Piaget, 1926; Vygotsky, 1932/1962). Social cognition refers to the fact that "cognitive development is both an individual and a social process" (Butterworth & Light, 1982, p. xiv). As Lave (1988, p. 1) explains, cognition is the "nexus of relations between the mind at work and the world in which it works." Rogoff and Lave (1984, p. 4) argue that cognitive activity takes place "in interaction with other people and use of socially provided tools and schemas[14] for solving problems. Cognitive activity is socially defined, interpreted, and supported." The "context provides information and resources that facilitate the appropriate solution of the problem at hand" (Rogoff & Lave, 1984, p. 4). Roloff and Berger (1982, p. 15) add that "In order to make sense of the complex information inputs involved in interaction, people are motivated to construct representations of reality." Social cognition is organized by means of scripts or conceptual schema, frameworks created by people to make sense of their environment and to communicate how they view that environment (Lave, 1988, p. 18).

One implication is that the study of cognition from inside the head must be shifted to the whole person "in action, acting with the settings of that activity" (Lave, 1988, p. 17). A second implication is that cognitive activities can no longer be assumed to be abstract and context-free.[15] Furthermore, knowledge and skills can not be easily transferred from one activity to another. Indeed, empirical studies indicate that heuristic rules upon which these generalizations are based often lead to significant errors in judgment. This is because differently organized learning experiences and tasks will yield different experiences and require that people develop different cognitive skills.

Cognition is socially distributed. Cicourel (1990, p. 223) writes that "The idea of socially distributed cognition refers to the fact that participants in collaborative work relationships are likely to vary in the knowledge they possess (Cicourel, 1974; Schutz, 1964) and must therefore engage each other in dialogues that allow them to pool resources and negotiate their differences to accomplish their tasks." Socially distributed cognition is

[14] Norman (1988a, p. 86) defines schemas as "knowledge structures that contain the general rules and information necessary for interpreting situations and for guiding behavior."

[15] This second implication strongly suggests that we must rethink and probably modify both our expectations about the transferability of what students learn and the pedagogical methods that we apply in the classroom.

analogous to distributed computing (Chandrasekaran, 1981; Gomez & Chandrasekaran, 1981; Smith & Davis, 1981). Cicourel (1990, p. 223) explains that "In automated systems, distributed problem solving refers to the cooperative solution of problems by a decentralized and loosely coupled collection of knowledge sources located in different processors." Smith and Davis (1981, p. 61) note that,

> The KS's [Knowledge Sources] *cooperate* in the sense that no one of them has sufficient information to solve the entire problem; mutual sharing of information is necessary to allow the group as a whole to produce an answer. By *decentralized* we mean that both control and data are logically and often geographically distributed; there is neither global control nor global data storage. Loosely coupled means that individual KS's spend the great percentage of their time in computation rather than communication.

This decentralization, due to decomposition of the complexity of the tasks, provides considerably increased efficiency because the subunits (Knowledge Sources) operate in parallel.[16] A decentralized approach to the organization of problem solving reduces uncertainty.[17] Resource constraints are also reduced by the distributed system because it is "highly modularized and offers conceptual clarity and simplicity of design" (Smith & Davis, 1981, p. 61).

Cooperation may come in many forms, but this article concentrates on sharing of information, or, as Smith and Davis (1981, p. 61) refer to it, "result-sharing," whereby, from time to time, "experts periodically report to each other the partial results they have obtained during the execution of individual tasks." The principles of feedback and control, leading to improving the quality of data and modifying the scientific design of panel surveys, for example, obviously require "task sharing," as well.

The form of a distributed system cannot be specified in advance because it will depend on the complexity of the task, level of uncertainty, and resource constraints. Fox (1981, p. 70) notes that "The major problem with designing distributed systems is deciding how the task should be decomposed and the [type of] control regime to be used, and this choice of organization is determined by features of the task (domain) and some measurement criteria [transaction analysis]." The laboratory and field research on communication networks suggest how to optimize communications among nodes. Design strategies in Part Three incorporate the results of these studies.

### *Summarizing the Argument: Core Concepts and Relationships, and Propositions Linking Communications, Social Knowledge and Error*

Figure 1 depicts the key flows and relationships that affect the data production, data utilization, and error production. Errors are embedded in the data production and utilization, which are eventually realized by data producers and users operating as experts, novices, or technological gatekeepers. Social and individual cognition facilitates the development of contextually-dependent conceptualizations, in the form of cognitive constructs and categories of meaning at the collective or group and individual levels. The realization of error in data production and utilization affects these cognitive processes in two ways: 1) Through individual and social communications, and 2) through the interplay of cognitive and communicative competencies operating within and between the individual and collective or group levels.

The lower half of Figure 1 reflects processes that operate at the *individual level* of cognition and communications. These include, as shown by the inner ring, shorter term, more episodic, and situationally-specific communications, languaging acts, and performance of tasks in which individual cognition operates as a mediating factor. The outer ring represents the accumulation of *individual* knowledge, experiences, and skills. Such accumulations define a contextual envelope in which individual understandings about data production and use develops through individual learning.

The upper half of Figure 1 describes processes that operate at the *social or collective level* of cognition and communications. This includes, as shown by the inner ring, shorter term, more episodic, and situationally-specific communications, functioning, and performance by collectives or groups. It focuses attention on social and technical issues in data production and utilization where social cognition serves as a mediating factor. The outer ring concerns the accumulation of socially and technically relevant knowledge, experiences, and skills by *collectives* or *groups*. Such accumulations define a contextual envelope for collective or group understandings about the data production and data utilization processes where organizational and situational contexts affect and are affected by social learning.

### *Three Propositions*

Table 1 displays the propositions. The first assumption (A1) is that error in the production and use of data

[16] Complexity, as used in this context, derives from work by Herbert Simon on "bounded rationality." Complexity is defined as "excessive demands on rationality. That is, the task requirements exceed current bounds on computational capacity" (Fox, 1981, p. 75). Our article is concerned with three aspects of complexity: Information, task, and communication.

[17] Uncertainty is defined as "the difference between information available and the information necessary to make the best decision" (Fox, 1981, p. 75). Fox adds three additional manifestations of uncertainty: Correctness of the information, knowledge lacking about possible outcomes of the decision, prediction of future states, and behavioral uncertainty.

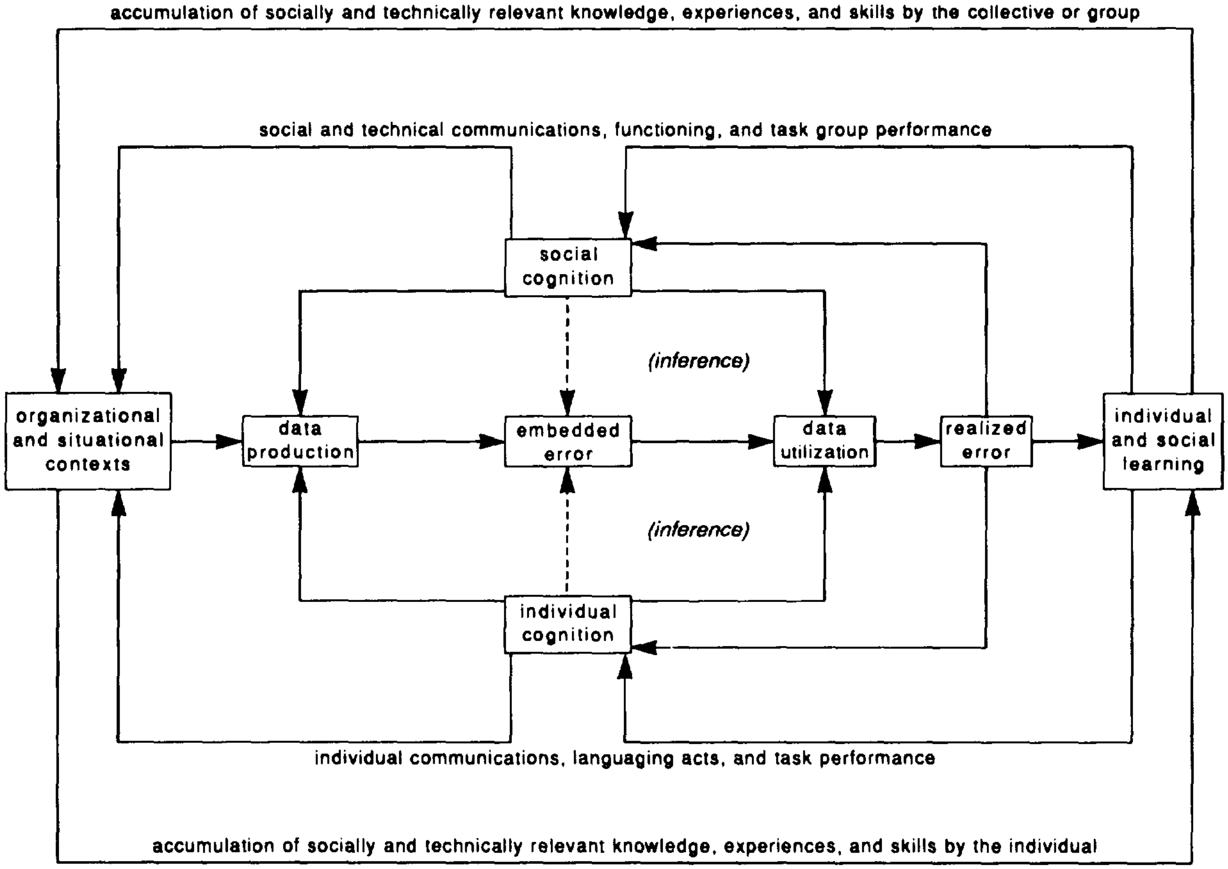
accumulation of socially and technically relevant knowledge, experiences, and skills by the collective or group
social and technical communications, functioning, and task group performance
social cognition
(inference)
organizational and situational contexts
data production
embedded error
data utilization
realized error
individual and social learning
(inference)
individual cognition
individual communications, languaging acts, and task performance
accumulation of socially and technically relevant knowledge, experiences, and skills by the individual

FIG. 1. Key flows and relationships affecting data production, data utilization, and error production processes.

is inevitable. Much of how people know and what people know involves learning on the job under conditions of substantial ambiguity, persistent uncertainty, and significant inconsistencies, gaps, and idiosyncratic quirks in their knowledge and understanding over what to do next with the data at hand. Trial and error learning is the usual way, and not the exceptional circumstance, that people utilize to develop an understanding about data and how they can be or are used.

The first proposition follows from this first assumption.

(P1.1) The greater the degree of ambiguity and uncertainty in social and situational contexts for which data are needed, desired or required;
(P1.2) The greater the likelihood that inconsistencies, gaps and quirks exist in the distribution of knowledge and understandings about data, its production and use;
(P1.3) The greater the extent to which trial and error learning dominates the production and use of data;
(P1.4) The greater the extent to which errors occurring in the production and use of data will be socially produced (rather than generated solely by individual mistakes in judgment);
(C1.1) Therefore, the greater the probability that errors will remain undetected in the production and use of data.

The second assumption (A2) is that data production and use involve communication processes that take place in a variety of social and organizational settings and through processes of individual and social cognition. Language serves a central function in communicating ideas and information about data, especially complex data. Language itself introduces variation in interpretive schemas and in communication of those schemas. Data represent a conversion of language forms from discourse to some type of symbolic representation (text, numbers, pictures/displays). The conversion process, however, introduces variation in the representation, meaning, and value attached to data.

A variety of communication network structures are possible to communicate information about data. Some network forms are more effective and efficient, and produce greater participant satisfaction than others, depending on the nature of the communication task regarding data production and data use. The occurrence of error in data depends upon the phenomenology of language, the interpretive processes associated with languaging as an act, the distribution of so-

TABLE 1. Assumptions, propositions, and conclusions.

| | | | |
|---|---|---|---|
| PROPOSITION SET #1 | | | |
| Assumption 1: | The occurrence of errors is inevitable in the production and use of data. | | |
| | Propositions | | Conclusions |
| P1.1 | The greater the degree of ambiguity and uncertainty in social and situational contexts for which data are needed, desired, or required | C1.1 | The greater the probability that errors will remain undetected in the production and use of data. |
| P1.2 | The greater the likelihood that inconsistencies, gaps, and quirks exist in the distribution of knowledge and understandings about data | | |
| P1.3 | The greater the extent to which trial and error learning dominates the production and use of data | | |
| P1.4 | The greater the extent to which errors are socially produced | | |
| PROPOSITION SET #2 | | | |
| Assumption 2: | Data production and use involve communication processes that take place in a variety of social and organizational settings and through processes of individual and social cognition. | | |
| | Propositions | | Conclusions |
| P2.1 | The greater the complexity of social and organizational communication structures and processes | C2.1 | The greater the probability that errors will significantly impact on and remain undetected in the production and use of data. |
| P2.2 | The greater the cognitive load and degree of cognitive complexity experienced by data producers and users | | |
| P2.3 | The greater the likelihood that communication network failures and breakdowns occur | | |
| PROPOSITION SET #3 | | | |
| Assumption 3: | Knowledge is socially produced and derived from experiences with events, situations, activities, objects, people, and use of data that describe these elements | | |
| | Propositions | | Conclusions |
| P3.1 | The greater the extent of variability in the distribution of knowledge about data production and uses between novices, experts, and technological gatekeepers | C3.1 | The greater the probability that errors will remain undetected in the production and use of data. |
| P3.2 | The extent to which deviations in the structures, processes, and content of communications about data production and uses remain ignored or are out of alignment between experts, novices, and technological gatekeepers. | | |

cial cognition, and the dynamics of social interaction in one or more social networks.

The second proposition follows from this second assumption.

(P2.1) The greater the complexity of social and organizational communication structures and processes;

(P2.2) The greater the cognitive load and degree of cognitive complexity experienced by data producers and users;

(P2.3) The greater the likelihood that communication network failures and breakdowns occur;

(C2.1) Therefore, the greater the probability that errors will significantly impact on and will remain undetected in the production and use of data.

The third assumption (A3) is that knowledge is socially produced and derived from experiences with events, situations, activities, objects, people, and the use of data that describe these elements. Discovery, diagnosis, correction, control, and prevention of error in data depend upon the social distribution of knowledge, which is created through relationships between experts, novices, and technological gatekeepers. Knowledge processes (learning, remembering, and transmitting information) are unevenly distributed within any social or organizational setting and lead to great variations in individual and collective understandings about the production and use of data. This occurs because knowledge is produced in a variety of venues (e.g., work situations, tacit and explicit observations, instructions, guided participations, etc.). Once created, this knowledge is transmitted over a variety of social and organizational communication network structures. This means that decisions about the production and use of data are socially distributed across participants and through time. This leads to cumulative or contradictory decisions on ways of transferring knowledge from experts to novices. The technological boundary spanner or gatekeeper can support this vital function by facilitating communication of information about data, its production, and its uses between experts and novices.

The third proposition then is that:

(P3.1) The greater the extent of variability in the distribution of knowledge about data production and uses between novices, experts, and technological gatekeepers;

(P3.2) The greater the extent to which deviations in the structures, processes, and the content of communications about data production and uses remain ignored or are "out of alignment" between experts, novices, and technological gatekeepers;

(C3.1) Therefore, the greater the probability that errors will remain undetected in the production and use of data.

## Part Three. Designing for Error: User-Centered Design

Production of error results from communication failures. Knowledge, information, and data regarding what is known about complex data are not communicated from one individual to another, between groups, or organizations and organizational subunits. Developing a distributed task system of social cognition requires effective communication. There are differences in the extent of knowledge and the nature of assumptions about what is known about the data by whom. A distributed task system for producing and using complex data must be designed to transfer information, reduce knowledge differences, communicate background assumptions underlying the data, and create a permanent repository of knowledge, experience, and skills.

According to Norman (1988b, p. 131), if a user makes an error, there is probably a good reason for it. If an analyst makes a mistake, it was probably because the information was unavailable or misleading. If the analyst did one thing, but intended to do another (a "slip"), it is probably due to a fault in the design of the task or information production system. Although it is not possible to design systems so that users make no errors, systems should be designed to: 1) Understand the causes of error and design to minimize those causes (see also Lewis & Norman, 1986); 2) make it possible to reverse actions—to "undo" them—or make it harder to do what cannot be reversed; 3) make it easier to discover the errors that do occur, and make them easier to correct; 4) change the attitude of those who work with data toward learning from errors rather than punishing them when errors occur; and, 5) reconceptualize the management of error in terms of opportunities to learn rather than situations of demonstrated incompetence.

Improvements must be made in conceptualizations of how people work with data: To view those people who produce and utilize data as people who perform tasks that generate some desired result from a series of imperfect approximations. People who work with data should not be viewed as people who either avoid or commit errors. Instead, their actions must be understood as a series of approximations toward a desired result.

To meet these design goals, information systems must function to support, in fundamental ways, the development of two interdependent competencies. Both *communicative competence* and *cognitive competence* must be developed to produce and utilize complex data. Several key strategies and principles are presented below that will move toward systems designs and operations which support these two competencies.

### *Designing for Communicative Competence*

These strategies have been strongly influenced by theorists who view the design of information systems as "tools for conversation." Conversation takes place in a social setting. Conversation is necessary to avoid breakdowns which are always on the verge of occurring, although, as Winograd and Flores (1987, p. 158) note, "It is impossible to completely avoid breakdowns by design, since it is in the nature of any design process that it must select a finite set of anticipations from the situation." However, developers can "partially anticipate situations where breakdowns are likely to occur (by noting their recurrence) and provide people with the tools and procedures they need to cope with them."

According to Winograd and Flores (1987, p. 162), communicative competence "means the capacity to express one's intentions and take responsibilities in the networks of commitments that utterances and their interpretations bring to the world." Communicative competence requires that systems developed include some capability to learn about the "fundamental relationships between language and successful action." Therefore, systems must be designed to focus on the use of language to characterize data. This requires a different view of language that extends beyond simply description. Systems must be designed to recognize that language is a form of action which creates commitments, in the form of promises and intentions to perform some action in the future. According to this view, language serves as the generative force in creating and changing our understanding about the social contexts in which we operate on an everyday basis.

To improve communicative competence in the domain of data production systems, opportunities for online, real-time learning through collective social interaction must be developed. This, of course, argues for a different approach to education than simply the formal classroom experience. Designers must seek to develop online learning systems that facilitate deeper and wider understanding of the linkages between everyday language and successful practical action involving large scale, complex data production systems. This "practical" perspective argues for communication systems that can reveal and make explicit those meanings and interpretations about the contents and procedures associated with the production and utilization of data. Such improvements are essential for large-scale, complex data systems because these meanings, once established, are rarely questioned.

Yet it is these implicit everyday understandings and interpretations about what the data mean and how they are and can be used that serve as the basis for breakdowns. Such breakdowns are represented by slips, mistakes, and miscommunication about the meaning of data which are revealed at various points in the data production and utilization life cycle. Such breakdowns represent opportunities for learning about data and for improving the quality of data, and for enhancing data production and data utilization systems. Rather than avoid error it must be embraced, especially, when it is realized through breakdowns in data production and use. Embracing error in this fashion allows greater control to be exercised when error does occur but would otherwise remain concealed (which is, of course, the worst error of all).

*1. Create a permanent repository of conversations about error.* Much of what is known about data is neither documented or communicated. It is permanently lost. An expanded view of information systems requires an integrated conceptualization about data, its production, and its use. Such designs incorporate expertise and knowledge of multiple data producers and users and take advantage of social interactions that are a natural part of most work. A distributed system of social cognition must create and communicate an audit trail of conversations about what is known, understood, and assumed about complex data. Such an information system must also include a variety of analytical functions to sift through the conversations and, eventually, develop a "global view" of knowledge about a particular set of complex data.

Because the database designer "designs the language that creates the world in which the user operates" (Winograd & Flores, 1987, p. 165), it is essential that database designs include the particular background assumptions, presuppositions, and cultural values of the designer, producer, and user. The cultural biases of designers, producers, and users must be brought to the attention of each participant in the entire system of data production and use. This requires information systems designs capable of creating a self-documenting history or system "autobiography" of data production and use. A system "autobiography" (i.e., documentation of formal policies, procedures and rules for producing and using data) would also track the network of interactions between producers and users. This type of design forces both data producer and user to communicate and exchange information about their operative, and often informal, policies and practices. Put simply, knowledge held by data producers and users must be communicated, transferred, and shared on a continuing basis if they are to develop information systems capable of learning about and controlling error.

*2. Prototype the data production and data utilization process.* Prototyping maximizes opportunities for learning about information system specifications (e.g., system needs, requirements, and uses) when prevailing conditions can be characterized by a high degree of complexity, a significant amount of irreducible uncertainty, and a lack of identifiable structure to guide system development efforts (Davis & Olson, 1985, pp. 566–568). A prototyping approach encourages the development of information system specifications through a process of iterative design and experimentation with limited "models" of input, processing, and output functions that the system is expected to perform (Davis & Olson, 1985, pp. 566–568). This design strategy relies on the active participation of end-users to provide knowledge and experience as an aid in determining system specifications. It offers an opportunity for the system designer to learn about an end-user's mental model of how tasks will be performed and the conditions under which end-users can and will transform data into something "of value" (i.e., information). It requires that end-users actively participate in developing and reacting to design changes. Prototyping allows end-users to learn how the system can satisfy their information needs while helping them to perform required tasks. Because prototyping relies on an iterative process of analysis, design, and implementation, it is more efficient than traditional approaches, enabling designers to construct less costly, but highly limited, systems to see how users will or will not use it (Boar, 1986). It also enables changes to occur in the information system specifications or designs before commitments become irrevocable (Rubin, 1986).

Most importantly, the prototyping strategy facilitates dialogue between information system users and system designers and allows the designers to modify quickly specifications in light of new information and learning (Martin, 1982). Gould and colleagues (1991, p. 75) contend that evidence exists to confirm that the user focus creates a "usability design process [that] leads to systems, applications, and products that are easy to learn, contain the right functions, are well liked, and safe." Evidence from other studies also indicates that participation in system design efforts increases user "literacy" and satisfaction with the system (Baronas & Reis, 1988; Baroudi, Olson, & Ives, 1986; Hirschheim, 1985; Montezemi, 1988; Norman, 1988a, 1988b). Prototyping is most useful when it supports an active learning process about the system needs, requirements, and uses.

Prototyping can be employed to model data collection, processing, and utilization efforts when the system is first being created or any time a major policy or procedural change occurs. This form of prototyping models the heuristic rules and actions taken by those people involved at key points in the data production and utilization process. When linked with the communication strategy cited above (Recommendation #1), prototyping provides an early warning of those data production and utilization activities that are likely to cause the occurrence of one or more errors. Early detection of error is important since, as noted by Lewis and Norman (1986, p. 419), early detection is "the first step toward recovery" and toward proper system functioning, as well.

*3. Create a communication network for shared conversations.* Underlying a cooperative task system is "a central assumption . . . that the knowledge sources must cooperate to solve a problem because no one source has enough information to do the job" (Cicourel, 1990, p. 223). Successful communicative acts depend on shared knowledge. "Communication with human beings," writes Mark (1986, p. 219), "requires shared understanding of the way the world works."

Management of complex data is not only managing data, but also about managing the coordination of information needs and knowledge about the data. A common framework for coordinating information is necessary if analysts are to "accomplish a cognitive performance" (Rogoff & Gardner, 1984, p. 97). This common framework for coordinating information must be established by the information system designer working in concert with data producers and users. Structuring dialogues and conversation about complex data requires the creation of formal roles to coordinate the flow of information and communications about complex data. This role is defined here as the "information gatekeeper."

Communication/information systems need to be designed to make information both easily available and accessible (Culnan, 1985). Information needs to be organized for ease of extraction and available when needed.[18] Both criteria need to be addressed not only when designing for communicative competence, but also when designing for cognitive competence.

### *Designing for Cognitive Competence*

Two strategies for designing for cognitive competence are emphasized: Achieve an understanding of the data and create a learning environment for the detection and correction of error.

*1. Provide a good mental model of the data.* Norman (1988b, p. 70) writes that "Good mental models help people derive appropriate behavior for situations that are not remembered or never before encountered." A good mental model makes it easier for people to learn and interpret information, and, consequently, detect error. Brown (1986, p. 466) writes about mental models as related to computer systems, but his comments are also relevant to the representation of complex data:

> Mental models of how the system functions relative to both its constituent parts and a given task provide the most stable and robust basis for understanding. Such models are also a crucial resource for facilitating informal discovery, learning through a sort of task-oriented empiricism, as they [users] form a cognitive structure about which hypotheses can be formed and tested.

A robust understanding of the data and how they are organized means that users will have a greater sense of control over the system, and this will lead to improvements in detecting and recovering from error.

Conceptual models of the data, especially competing conceptual models, must be made more explicit. Norman (1986) suggests that conceptual models provide a scaffolding upon which to build bridges across what he calls "the gulf between execution and evaluation." The scientific design and procedures constitute the conceptual model of the data. This would allow inference from

[18] Tulving and Pearlstone (1966) and Fischhoff et al. (1987) distinguish between the concepts of availability and accessibility, and suggest that the quality of an informational resource can be measured by these two criteria (p. 33). The availability criterion refers to the "amount of information that it [the resource] contains." The accessibility criterion refers to the "ease with which users extract information from it [the resource]."

data that requires interpretation to be consistent with the design and procedures used to execute the design.

The conceptual model of the data means that data must have an explicit logical structure in order to reduce cognitive complexity and ambiguity (see David, 1991; David & Robbin, 1990a). This requires that objects, properties, contexts, and relationships must be explicitly defined and systematically organized, so that people can work effectively and make inferences about the next steps to take in problem-solving (Reichman, 1986, p. 310). For complex data, this implies that the entities and the relationships must be clearly displayed by semantic principles that organize each part of a database. In this regard three aspects must be considered: 1) The object to which an item in an instrument refers (referent); 2) the bounding reference period; and 3) the attribute that is being elicited. The aggregation of entities must be clearly identified, and the aggregation of measures over several entities and the attribution of aggregates to those entities must be clearly described. Repeated measures must be identified and unambiguously labeled. Measures taken at different points in time that are not identical must be clearly identified to reduce ambiguity and uncertainty about what the data and data interrelationships really represent. Precise relationships describing their similarities can then be constructed. Finally, the timing and length of events must be clearly identified, including intervals in which respondents provide or fail to provide adequate responses.

*2. Create a supported learning environment to encourage the self-detection of error.* Although considerable prior knowledge is required to use complex data effectively, formal instruction can rarely, if ever, provide all the skills needed. Knowledge of how to use a data set appropriately is usually acquired "on the job" during actual use. As such, errors will always occur during a training period while users familiarize themselves with the data. Furthermore, knowledge of the data will be differentially distributed among experts and novices according to their career trajectories, experiences, knowledge, and acquired skills.

Siefert and Hutchins (1992, p. 17) have observed that four elements provide an opportunity to detect error: Access, knowledge and expectation, attention, and perspective. Access means the "ability to observe errors being made." Knowledge and expectation concern the "ability to judge that the process or product will produce error." Attention relates to the "monitoring of activity." Differences in perspective pertain to "noting of discrepancies from a different point of view." These elements have implications for the organization of learning opportunities and the role of expert as a teaching coach (i.e., a transmitter of knowledge, skills, and values or culture) in a distributed system of social cognition. These strategies emphasize information systems as a learning environment, cognition as a generative process, and the interdependency of relations between, and active involvement of, experts and novices.

Analysts are limited by how much information they can retrieve and evaluate. Because there are severe limitations on memory and attention, systems must be designed to reduce the number and structure of tasks that the user is required to perform on the data. It is therefore essential to simplify what must be remembered. Norman (1988b) has a number of suggestions that are relevant for reducing the cognitive complexity of tasks associated with longitudinal panel studies. He first suggests the need to minimize planning or problem solving on the part of the user while maximizing experimentation and feedback. This implies, for example, the use of a sample database that lowers the high costs of learning through quick retrieval of results and the development of ad hoc queries. Norman's second suggestion is to modularize tasks, so that each module has limited properties and limited features. By modularizing what must be analyzed by the user, the system designer reduces the amount of data needed to be understood at any one time. The effect of modularizing is to make it easier to discover errors that do occur and easier to correct them. Making the entity-attribute relationships more visible, for example, will "help the user identify problems" with the data, [and] it can also help the system identify how it can help the user" (O'Malley, 1986, p. 388). Norman's third suggestion is to allow only those features of the data that are absolutely necessary. This implies, for example, that database designers make available the measurements as recorded by the instrument and minimize the number of "constructed" variables. This approach, in effect, encourages "filtering" to reduce memory load. The fourth suggestion made by Norman is to provide mental aids that form a knowledge base. Lewis (1986, p. 183) suggests that certain mental aids should be designed to assist information selection because users "often make errors by assuming that actually irrelevant information is related to their concerns of the moment."

At least three applications derive from Lewis's suggestion. First, metadata should be developed to make the background assumptions and contexts of the data explicit, explicate the meaning of the data, "form principles on which to build understanding" (Mark, 1986, p. 220), show alternative courses of action, and help users evaluate the effects of their actions. Systems should be designed with extended browsing capabilities that familiarize users with the categories and underlying conceptual structure of both the information system and the database (Cleal, 1988, p. 230; Fischhoff, MacGregor, & Blackshaw, 1987). Help systems should be employed because they can extend the user's memory (David & Robbin, 1992; Norman & Draper, 1986, p. 356; Owen, 1986) and because they more closely resemble the distributed model of information retrieval by which understanding and learning actually takes place (diSessa, 1986).

The task system creates a supported learning situation whose objectives are to extend the analyst's current skills and knowledge to a higher level of competence, to build new skills into the analyst's existing knowledge structure (Greenfield, 1984; Rogoff & Gardner, 1984), and to increase the analyst's sense of control over the data and information system. Fundamental to ensuring success in learning is the application of principles associated with "scaffolded" learning environments. Through "guided participation" an expert transfers information, guides the transfer of knowledge skills from other problem solving contexts, and evaluates the actions of the novice. The expert monitors the process that novices follow and assesses their performance. Experts help novices move toward independence by formulating questions that enable novices to eventually discover answers on their own (O'Malley, 1986). An expert plays an essential role in structuring the novice's interaction with data. Expert diagnosis of errors made by novices promotes error correction. As Seifert and Hutchins (1992, p. 23) note, "Beyond correcting an error that has occurred, a consequence of having engaged in diagnosing the cause of an error may be a new insight about the task processes." Feedback is provided through demonstration, instruction, and question-asking and -answering to "guide the discovery of concepts underlying the solution" (Seifert & Hutchins, 1992, p. 26).

Two principles are emphasized with this strategy: Exploration, which may occur through hypothesis-testing, for example, and error-based learning, which may arise by making mistakes, asking questions, seeking advice, or searching for information. Both activities are carried out through a dialogue between novice and expert. In other words, as O'Malley (1986, pp. 397–398) argues, "It is not enough to produce only the information alone—the system should also support question discovery in order to support the 'whole user' activity of information acquisition—of users helping themselves" to be able to ask the question in the first place (see LaFrance, 1992; Graesser, Person, & Huber, 1992).

Underlying the strategy of designing an information system as a learning environment is the principle of a cooperative task environment. People depend on each other for information, advice, and problem solving. Implied by this strategy is the design of an information system that develops and relies on local experts and a sense of community which is created when data producers and users work together on the same data set (cf., Bannon, 1986). Participation creates a redundancy of knowledge and experience as the number of experts increases. The distribution of knowledge among novices and experts ensures that most errors will be detected through the multiple perspectives and communication among the participants.

## Concluding Remarks

A very large investment has been made in the production and use of public data from administrative records and longitudinal panel surveys, but subsequent use of these complex data has been much smaller than anticipated. It is also important, however, to acknowledge that conditions other than the design of appropriate information systems and services precluded using large-scale, complex datasets until very recently. These other conditions include: 1) A significant intellectual and capital investment required because of the significant size, scope, and complexity of large-scale datasets; 2) the lack of low-cost computer-based technologies for efficient and low-cost data reorganization and retrieval, communication of scientific information, and exchange of data; and, 3) the recent emergence of a national information and social science infrastructure for conducting research and public policy analysis which, nevertheless, was not designed to respond quickly to a dynamic environment for data production and utilization (David & Robbin, 1992). It should also be noted that the mental models and prevailing paradigms which shape social science activity—that provide frameworks guiding the cultural values and activities about the enterprise itself—remained unchanged. By and large, the techniques and tools have not been connected to the process of scientific discovery and to improvements in theory, measurement, and data quality (Hall, 1992). One consequence has been little if any interest in applying advances in information technology to improve the quality of complex data and to enhance the performance and results obtained from the current social science infrastructure.

At the same time, however, developments in concepts and methods in information systems designs have, for the most part, proceeded without an adequate theoretical foundation. Technical systems continue to be built in the United States without much attention to the social systems in which they are embedded—this despite continuing acknowledgment and diagnosis of the fundamental problem. Collaboration between social scientists, information scientists, and computer scientists has been minimal if nearly non-existent (Robbin, 1995). And this, in itself, may explain why many technological innovations in the social sciences have failed, and why the current social science infrastructure in the United States has not evolved in response to these new technologies.

This article makes the case that the quality of data produced and used can be improved if the design of information systems rests on a theoretical foundation of social science theory and practice. Attention has been drawn to a large body of theory and research that can be used to design information systems that help us discover, understand, and control error which occurs through a variety of interacting organizational, social, and cognitive processes. A change in perspective in how professionals approach and act toward the process of designing and using information systems is necessary. A definitive answer regarding the design of organizations and information systems cannot be given, unfortunately—it is a highly contingent endeavor. The scale, structure, and technologies embedded in organizations that produce and use complex data will vary according to the type of data, target clientele, and purposes which the data serve. Furthermore, organizational theory does not allow someone to predict with any certainty what the likely consequences or effects of particular types of organizational designs might be on the quality of the data, or whether they are produced and used with greater effectiveness or efficiency. It is therefore impossible to predict or prescribe "optimal" organizational designs that would consistently support the efficient and effective production and use of complex data.

Nevertheless, the philosophic stance and commitment must be to design information systems that more closely reflect how people actually interact in social situations to ask questions, make discoveries, to learn, and to solve problems. The approach discussed in this article for designing information systems that produce and utilize complex data requires a "distributed system of social cognition"—one that reflects how the social science enterprise is actually conducted. As such, this perspective offers the potential for integrating a wider variety of technological innovations to improve the enterprise of social science inquiry. The authors' hope is that this discussion will foster the necessary conversations that must occur between members of those communities who produce data, design information systems, and utilize data. In this way, members of all communities can begin to more fully develop systems, policies, procedures, and rules that will significantly improve the production and utilization of complex data.

## References

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

Bannon, L. J. (1986). Helping users help each other. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 399-410). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baronas, A.-M. K., & Reis, M. (1988, March). Restoring a sense of control during implementation: How user involvement leads to system acceptance. *MIS Quarterly, 13,* 111-124.

Baroudi, J. J., Olson, M. H., & Ives, B. (1986). An empirical study of the impact of user involvement on system usage and information satisfaction. *Communications of the ACM, 29,* 232-238.

Boar, B. (1986, February). Application prototyping: A life cycle perspective. *Journal of Systems Management,* 25-31.

Bateson, A. G., Alexander, R. A., & Murphy, M. D. (1987). Cognitive processing differences between novice and expert computer programmers. *International Journal of Man-Machine Studies, 26,* 649-660.

Bavalas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America, 22*(6), 725-730.

Bodenhausen, G. V., & Wyer, R. S. (1987). Social cognition and social reality: Information acquisition and use in the laboratory and the real world. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 8-41). New York: Springer-Verlag.

Brittain, J. M. (1970). *Information and its users.* Bath, UK: Bath University Press.

Brown, J. S. (1986). From cognitive to social ergonomics and beyond. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 457-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Butterworth, G., & Light, P. (Eds.). (1982). *Social cognition: Studies of the development of understanding.* Chicago: University of Chicago Press.

Carlson, P. A. (1990, April). Square books and round books: Cognitive implications of hypertext. *Academic Computing,* 16-31.

Chandrasekaran, B. (1981, January). Natural and social system metaphors for distributed problem solving: Introduction to the issue. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11*(1), 1-5.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Cicourel, A. V. (1974). *Cognitive sociology: Language and meaning in social interaction.* New York: Free Press.

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

Clark, H. H., & Carlson, T. B. (1982). Speech acts and hearers' beliefs. In N. V. Smith (Ed.), *Mutual knowledge* (pp. 1-36). New York: Academic Press.

Clark, N. (1986). Tables and graphs as language. *Proceedings of the 18th symposium on the interface,* Fort Collins, CO (pp. 83-89). Washington, DC: American Statistical Association.

Cleal, D. M. (1988). ISSUE—a case study in expert system interfaces. In N. Heaton & M. Sinclaim (Eds.), *Designing end-user interfaces. State of the Art Report* (pp. 3-16). *15:8.* Maidenhead, Berkshire, UK: Pergamon Infotech Limited.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Comer, D. (1983). The computer science research network CSNET: A history and status report. *Communications of the ACM, 26*(8), 747-753.

Cotter, H. (1988). Birth of a network: A history of BITNET. *CUNY/ University of Computer Center Communications, 14,* 1-10.

Culnan, M. J. (1985). The dimensions of perceived accessibility to information: Implications for the delivery of information systems and services. *Journal of the American Society for Information Science, 36*(5), 302-308.

David, M. H. (1980). Access to data: The frustration and utopia of the researcher. *Review of Public Data Use, 8,* 327-337.

David, M. H. (1991). The science of data sharing: Documentation. In J. E. Sieber (Ed.), *Sharing social science data: Advantages and challenges* (pp. 91-115). Newbury Park, CA: Sage Publications.

David, M. H., & Robbin, A. (1990a). Database design for large-scale, complex data. *Proceedings of the 21st Symposium on the Interface between Statistics and the Computer,* April 1989, Orlando, FL. Alexandria, VA: American Statistical Association.

David, M. H., & Robbin, A. (1990b). Computation using information systems for complex data. *Proceedings of the Conference on Advanced Computing in the Social Sciences,* April 1990, Williamsburg, VA. Oak Ridge: Oak Ridge National Laboratory and Washington, DC: U.S. Bureau of the Census.

David, M. H., & Robbin, A. (1992, February). *Designing new infrastructures for the social sciences.* Final report to the National Science Foundation on the SIPP ACCESS project. Madison, WI: University of Wisconsin.

Davis, G., & Olson, M. (1985). *Management information systems: Conceptual foundations, structure, and development* (2nd ed.). New York: McGraw Hill.

Dervin, B. (1977). Useful theory for librarianship: Communication, not information. *Drexel Library Quarterly, 13,* 16-32.

Dervin, B. (1983). Information as a user construct: The relevance of perceived information needs to synthesis and interpretation. In S. A. Ward and L. J. Reed (Eds.), *Knowledge structure and the implications for synthesis and interpretation* (pp. 155–183). Philadelphia: Temple University Press.

Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In J. D. Glazier & R. R. Powell (Eds.), *Qualitative research in information management*. Englewood, CO: Libraries Unlimited.

Dervin, B., & Nilan, M. S. (1986). Information needs and uses. *Annual Review of Information Science and Technology, 21,* 3–33.

diSessa, A. A. (1986). Models of computation. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 201–218). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dolby, J. L., Clark, N., & Rogers, W. H. (1986). The language of data; A general theory of data. *Proceedings of the 18th Symposium on the Interface* (pp. 96–103). Washington, DC: American Statistical Association.

Eradi, Y. M., & Utterback, J. M. (1984). The effects of communication on technological innovation. *Management Science, 25,* 572–585.

Finholt, T., & Sproull, L. S. (1990). Electronic groups at work. *Organization Science, 1*(1), 41–64.

Finholt, T., Sproull, L., & Kiesler, S. (1990). In J. Galegher, R. E. Kraut, & C. Egido (Eds.), *Intellectual teamwork: Social and technological foundations of cooperative work* (pp. 291–325). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fischhoff, B., MacGregor, D., & Blackshaw, L. (1987). Creating categories for databases. *International Journal of Man-Machine Studies, 27,* 33–62.

Fox, M. S. (1981). An organizational view of distributed systems. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11*(1), 70–80.

Garfinkel, I., Corbett, T. J., MacDonald, M., McLanahan, S., Robbins, P. I., Schaeffer, N. C., & Seltzer, J. (1988, June). *Evaluation design for the Wisconsin Child Support Assurance Demonstration* (Report prepared for the Wisconsin Department of Health and Social Services). Madison, WI: Institute for Research on Poverty, University of Wisconsin-Madison.

Garnham, A. (1987). *Mental models as representations of discourse and text.* New York: John Wiley & Sons.

Garvey, W. D. (1979). Communication: The essence of science. New York: Pergamon Press.

Garvey, W. D., & Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and the creation of knowledge. *American Psychologist, 26*(4), 349–362.

Garvey, W. D., Lin, N., Nelson, C. E., & Tomita, K. (1972a). Research studies in patterns of scientific communication—I, General description of research program. *Information Storage Retrieval, 8,* 111–122.

Garvey, W. D., Lin, N., Nelson, C. E., & Tomita, K. (1972b). Research studies in patterns of scientific communication—II, The role of the national meeting in scientific and technical communication. *Information Storage Retrieval, 8,* 159–169.

Garvey, W. D., Lin, N., & Tomita, K. (1972c). Research studies in patterns of scientific communication—III, Information-exchange processes associated with the production of journal articles. *Information Storage & Retrieval, 8,* 207–221.

Garvey, W. D., Lin, N., & Tomita, K. (1972d). Research studies in patterns of scientific communication—IV, The continuity of dissemination of information by "productive scientists." *Information Storage Retrieval, 8,* 265–276.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teaming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99–130). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goleman, D. (1994, October 11). Peak performance: Why records fail. *The New York Times,* sec. C, pp. 1, 12.

Gomez, F., & Chandrasekaran, B. (1981). Knowledge organization and distribution for medical diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11*(1), 34–42.

Gould, J. D., Boies, S. J., Boies, & Lewis, C. (1991, January). Making usable, useful, productivity: Enhancing computer applications. *Communications of the ACM, 34,* 75–85.

Graesser, A., Person, N., Huber, J. (1992). Mechanisms that generate questions. In T. W. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems* (pp. 167–187). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Greenfield, P. M. (1984). A theory of the teacher in the learning activities of everyday life. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 117–138). Cambridge, MA: Harvard University Press.

Griffith, B. C., & Miller, A. J. (1970). Networks of informal communication among scientifically productive scientists. In C. E. Nelson & D. K. Pollack (Eds.) *Communication Among Scientists and Engineers* (pp. 125–140). Lexington, MA: Heath Lexington Books.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: John Wiley & Sons.

Guetzkow, H., & Simon, H. (1955). The impact of certain communication nets upon organization and performance in task-oriented groups. *Management Science, 1,* 233–250.

Hagstrom, W. D. (1965). The scientific community. New York: Basic Books.

Hagstrom, W. D. (1970). Factors related to the use of different modes of publishing research in four scientific fields. In C. E. Nelson & D. K. Pollock (Eds.), *Communication among scientists and engineers* (pp. 87–124). Lexington, MA: D.C. Heath.

Hall, S. S. (1992, 17 July). How technique is changing science. *Science, 257,* 344–349.

Hannaway, J. (1989). *Managers managing: The workings of an administrative system.* New York: Oxford University Press.

Hastorf, A. H., & Isen, A. M. (Eds.). (1982). *Cognitive social psychology.* New York: Elsevier North Holland.

Hesse, B. W., Sproull, L., Kiesler, S., & Walsh, J. P. (1993, August). Returns to science: Computer networks in oceanography. *Communications of the ACM, 36*(8), 90–101.

Hippler, H.-J., Schwarz, N., & Sudman, S. (Eds.). (1987). *Social information processing and survey methodology.* New York: Springer-Verlag.

Hirschheim, R. A. (1985, December). User experience with and assessment of participative systems design. *MIS Quarterly,* 295–304.

Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. *Organization Science, 2*(1), 88–115.

Hutchins, E. (1990, May). *Learning to navigate.* Unpublished manuscript, University of California, San Diego.

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

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment.* New York: Free Press.

Katz, R., & Tushman, M. (1979). Communication patterns, project performance, and task characteristics: An empirical evaluation and integration in an R&D setting. *Organizational Behavior and Human Performance, 23,* 139–162.

Krauss, R. M., & Fussell, S. R. (1990). Mutual knowledge and communicative effectiveness. In J. Galegher, R. E. Kraut, & C. Egido (Eds.), *Intellectual teamwork: Social and technological foundations of cooperative work* (pp. 111–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kraut, R. E., Egido, C., & Galegher, J. (1990). Patterns of contact and communication in scientific research collaborations. In J. Galegher, R. E. Kraut, & C. Egido (Eds.), *Intellectual teamwork: Social and technological foundations of cooperative work* (pp. 149–172). Hillsdale, NJ: Lawrence Erlbaum Associates.

LaFrance, M. (1992). Questioning knowledge acquisition. In T. W. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems* (pp. 11–28). Hillsdale, NJ: Lawrence Erlbaum Associates.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980, June). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life.* New York: Cambridge University Press.

Lewis, C. (1986). Understanding what's happening. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 171–185). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, C., & Norman, D. A. (1986). Designing for error. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 411–432). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lievrouw, L. A. (1988). Four programs of research in scientific communication. *Knowledge in Society, 1*(2), 6–22.

Linn, M. C., & Clancy, M. J. (1990, April). Designing instruction to take advantage of recent advances in understanding cognition. *Academic Computing,* 20–41.

March, J. G. (1991). Explorations & exploitation in organizational learning. *Organization Science, 2*(1), 71–87.

March, J. G., & Simon, H. A. (1958). *Organizations.* New York: John Wiley & Sons.

Mark, W. (1986). Knowledge-based interface design. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 219–238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Martin, J. (1982). *Application development without programmers.* Englewood Cliffs, NJ: Prentice-Hall.

McCall, L. (1989, April 20). DOR data documentation. Unpublished manuscript.

McGrath, J. E. (1984). *Groups: Interaction and performance.* Englewood Cliffs, NJ: Prentice-Hall.

McGrath, J. E. (1990). Time matters in groups. In J. Galegher, R. E. Kraut, & C. Egido (Eds.), *Intellectual teamwork: Social and technological foundations of cooperative work* (pp. 27–61). Hillsdale, NJ: Lawrence Erlbaum Associates.

McMillen, D. B. (1990, May). Redesigning the SIPP public use data files. *SIPP Supplement to the APDU Newsletter, 3*(2), 1.

McPhee, R. D., & Tompkins, P. K. (Eds.)(1985). *Organizational communication: Traditional themes and new directions.* Beverly Hills: Sage Publications.

Mehan, H. (1984). Institutional decision-making. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 41–66). Cambridge, MA: Harvard University Press.

Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of "unstructured" decision processes. *Administrative Science Quarterly, 21,* 236–275.

Mitroff, I. I., & Mason, R. O. (1981). *Challenging strategic planning assumptions.* New York: John Wiley & Sons.

Montezemi, A. R. (1988). Factors affecting information satisfaction in the context of the small business environment. *MIS Quarterly, 13,* 239–256.

Morris, R. C. (1994). Toward a user-centered information service. *Journal of the American Society for Information Science, 45*(1), 20–30.

National Academy of Sciences, National Academy of Engineering, Institute of Medicine, Committee on Science, Engineering, and Public Policy. (1989). *Information technology and the conduct of research: The user's view* (Report of the Panel on Information Technology and the Conduct of Research). Washington, DC: National Academy Press.

National Research Council, Committee on a national Collaboratory: Establishing the User-Developer Partnership. (1993a). *National collaboratories: Applying information technology for scientific research.* Washington, DC: National Academy Press.

National Research Council, Committee on National Statistics. (1993b). *The future of the survey of income and program participation.* Washington, DC: National Academy Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Norman, D. A. (1983a). Design rules based on analyses of human error. *Communications of the ACM, 4,* 254–258.

Norman, D. A. (1983b). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 7–14). Hillsdale, NJ: Lawrence Erlbaum Associates.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 31–62). Hillsdale, NJ: Lawrence Erlbaum Associates.

Norman, D. A. (1988a). Developments in computing and the user interface—emerging issues in end-user interface design. In N. Heaton & M. Sinclair (Eds.), *Designing end-user interfaces. State of the Art Report* (pp. 85–96). *15:8.* Maidenhead, Berkshire, UK: Pergamon Infotech Limited.

Norman, D. A., (1988b). *The psychology of everyday things.* New York: Basic Books, Inc.

Norman, D. A., & Draper, S. W. (Eds.). (1986). *User center system design: New perspectives on human-computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Olsen, R. J., David, M. H., & Sheets, C. T. (1991, May). The operation of SIPP and implications for the future. Unpublished manuscript.

O'Malley, C. E. (1986). Helping users help themselves. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 377–398). Hillsdale, NJ: Lawrence Erlbaum Associates.

Owen, D. (1986). Answers first, then questions. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 361–375). Hillsdale, NJ: Lawrence Erlbaum Associates.

Paisley, W. J. (1965). *The flow of (behavioral) science information: A review of the research literature.* Stanford, CA: Stanford University, Institute for Communication Research.

Paisley, W. J. (1972). The role of invisible colleges in scientific information transfer. *Educational Researcher, 1*(4), 5–19.

Paisley, W. J. (1984). Communication in the communication sciences. In B. Dervin, & M. Voight (Eds.), *Progress in communication sciences, 5,* 1–44. Norwood, NJ: Ablex Publishing Corporation.

Payne, H. W. (1980). Information processing theory: Some concepts and methods applied to decision research. In T. S. Wallstein (Ed.), *Cognitive processes in choice and decision behavior* (pp. 95–116). Hillsdale, NJ: Lawrence Erlbaum Associates.

Piaget, J. (1926). *The language and thought of the child.* New York: Harcourt, Brace.

Phillips, E., & Garfinkel, I. (1992, May). *Changes over time in the incomes of nonresident fathers in Wisconsin.* (DP #967-92). Madison, WI: Institute for Research on Poverty, University of Wisconsin-Madison.

Reichman, A. R. (1986). Communication paradigms for a window system. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 285–314). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reif, F. (1980). Theoretical and educational concerns with problem solving: Bridging the gaps with human cognitive engineering. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 39–52). Hillsdale, NJ: Lawrence Erlbaum Associates.

Robbin, A. (1992). Social scientists at work on electronic research networks. *Electronic Networking: Research, Applications and Policy, 2*(2), 6–30.

Robbin, A. (1995). SIPP ACCESS, an information system for complex data: A case study in creating a collaboratory for the social sciences. *Internet Research: Electronic Networking Applications and Policy, 5*(2), 37–66.

Roberts, J. (1964). The self-management of cultures. In W. Goodenough (Ed.), *Explorations in cultural anthropology: Essays in* *honor of George Peter Murdock* (pp. 433–454). New York: McGraw-Hill.

Rogers, E. M. (1994). *The diffusion of innovations* (4th ed.). New York: The Free Press.

Rogers, E. M., & Kincaid, D. L. (1981). *Communication networks: Toward a new paradigm for research.* New York: Free Press.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* New York: Oxford University Press.

Rogoff, B., & Gardner, W. (1984). Adult guidance of cognitive development. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 95–116). Cambridge, MA: Harvard University Press.

Rogoff, B., & Lave, J. (1984). *Everyday cognition: Its development in social context.* Cambridge, MA: Harvard University Press.

Roloff, M. E., & Berger, C. R. (1982). Social cognition and communication: An introduction. In M. E. Roloff & C. R. Berger (Eds.), *Social cognition and communication* (pp. 9–32). Beverly Hills, CA: Sage Publications.

Rubin, B. M. (1986). Information systems for public management: Design and implementation. *Public Administration Review (Special Issue)*, 540–552.

Ryscavage, P. (1987). *SIPP: Filling data gaps on the poverty and social welfare fronts.* (SIPP Working Papers No. 8705). Washington, DC: U.S. Bureau of the Census.

Saracevic, T., & Kantor, P. (1988a). The study of information seeking and retrieving. II. User, questions and effectiveness. *Journal of the American Society for Information Science, 39*(3), 177–196.

Saracevic, T., & Kantor, P. (1988b). The study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science 39*(3), 197–216.

Schatz, Bruce R. (1992). Building an electronic community system. *Journal of Management Information Systems, 8*(3), 87–107.

Schutz, A. (1962). *Collected papers I: The problem of social reality.* Vol. I. Introduction by M. Natanson, (Ed.), with a preface by H. L. Van Breda. The Hague: Martinus Nijhoff.

Schutz, A. (1964). *Collected papers II: Studies in social theory.* The Hague: Martinus Nijhoff.

Schutz, A. (1967). *Collected papers III: Studies in phenomenological philosophy.* I. Schutz, (Ed.), with an introduction by A. Gurwitsch. The Hague: Martinus Nijhoff.

Seifert, C. M., & Hutchins, E. L. (1989). Learning from error. *Proceedings of the 11th Annual Conference of the Cognitive Science Society,* August 16–19, 1989, Ann Arbor, MI, (pp. 42–49).

Seifert, C. M., & Hutchins, E. L. (1992). *Error as opportunity: Learning in a situated task.* Ann Arbor: University of Michigan and San Diego: University of California.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization.* New York: Doubleday/Currency.

Shaw, M. E. (1978). Communication networks. In L. Berkowitz (Ed.), *Group processes: Papers from advances in experimental social psychology* (pp. 313–349). New York: Academic Press.

Simon, H. A. (1979). *Models of thought.* New Haven: Yale University Press.

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

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325–348). Hillsdale, NJ: Lawrence Erlbaum Associates.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28,* 1–39.

Smith, R. G., & Davis, R. (1981). Frameworks for cooperation in distributed problem solving. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11*(1), 61–70.

Spitzberg, B. H. (1987). Issues in the study of communicative competence. In B. Dervin & M. J. Voigt (Eds.), *Progress in communication sciences, 8,* 1–46.

Steele, G. (1984). *COMMON LISP: The language.* Bedford, MA: Digital Press.

Sutherland, J. W. (1978). *Administrative decision-making: Extending the bounds of rationality.* New York: Van Nostrand Reinhold Company.

Taylor, R. S. (1968). Question negotiation and information seeking in libraries. *College & Research Libraries, 29,* 178–194.

Taylor, R. S. (1986). *Value-added processes in information systems.* Norwood, NJ: Ablex Publishing Corporation.

Taylor, R. S. (1991). Information use environments. In B. Dervin (Ed.), *Progress in communication sciences,* (Vol. 10, pp. 218–255). Norwood, NJ: Ablex Publication Corporation.

Taylor, R. W., & Utterback, J. M. (1975, May). A longitudinal study of communication in research: Technical and managerial influences. *IEEE Transactions on Engineering Management, EM-22,* 80–87.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381–391.

Tushman, M. L. (1977). Special boundary roles in the innovation process. *Administrative Science Quarterly, 22,* 587–605.

Tushman, M. L., & Katz, R. (1980). External communication and project performance: An investigation into the role of gatekeepers. *Management Science, 26*(11), 1071–1085.

Tversky, A., & Kahneman, D. (1974, September). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Tversky, A., & Kahneman, D. (Eds.). (1982). *Judgment under uncertainty.* New York: Cambridge University Press.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, M.I.T. Press. (Original work published 1932).

Webber, D. J. (Winter, 1991/92). The distribution and use of policy knowledge in the policy process. *Knowledge and Policy: The International Journal of Knowledge Transfer and Utilization, 4*(4), 6–35.

Wilensky, H. L. (1967). *Organizational intelligence: Knowledge and policy in government and industry.* New York: Basic Books.

Wilson, P. (1995). Unused relevant information in research and development. *Journal of the American Society for Information Science, 46*(1), 45–51.

Winograd, T., & Flores, F. (1987). *Understanding computers and cognition: A new foundation for design.* Reading, MA: Addison-Wesley Publishing Company.