# Time-Line Interviews and Inductive Content Analysis: Their Effectiveness for Exploring Cognitive Behaviors

**Linda Schamber**
*School of Library and Information Sciences, University of North Texas, P.O. Box 311068, Denton, TX 76203-1068*

**In studies of information users' cognitive behaviors, it is widely recognized that users' perceptions of their information problem situations play a major role. Time-line interviewing and inductive content analysis are two research methods that, used together, have proven extremely useful for exploring and describing users' perceptions in various situational contexts. This article describes advantages and disadvantages of the methods using examples from a study of users' criteria for evaluation in a multimedia context.**

## Introduction and Background

Naturalistic approaches to research and the use of qualitative methods have long been mainstream in the social sciences. Information scientists have contributed to the methodological toolbox by adapting existing methods and developing new methods for studying human information behavior. A distinct body of research has emphasized the role of information users' problem situations in influencing their information seeking and using behaviors. Two methods that have been quite successful at capturing users' cognitive perceptions in various situational contexts have been time-line interviewing and inductive content analysis. This article explains specific aspects of the methods, including advantages and disadvantages, using as an example a study of users' criteria for evaluation of information and information sources.

Time-line interviewing is a technique derived from anthropology, ethnography, and clinical psychology (among others). In information science, it is most often associated with sense-making theory as described by Dervin (1983, 1992, 1997). Sense making refers to the ways in which individuals perceive and bridge cognitive gaps in order to make personal sense of the world. Sense making assumes that individuals are constantly constructing and reconstructing personal views of internal and external reality based on their perceptions of their current situations, and that these views are a fundamental part of who they are at a given time. Although the philosophical underpinnings of sense-making theory are complex, subtle, and sometimes controversial, its holistic approach provides a useful framework in information science for observing information seeking from information users' perspectives. In many cognitively oriented user studies, the context is conceptualized as the situation that motivates the user to seek information, and the cognitive gap as the information need or problem.

The emphasis in this article, however, is not on sense making per se, but rather on two methods used together to discover and capture users' perceptions about their situations. In naturalistic research, the researchers are expected to collect the data in a real-world setting. If researchers take a grounded theory approach, or derive findings from observations of real-life phenomena, they are expected to avoid imposing a priori structures on the data. When the data are cognitive perceptions, the researcher is particularly challenged to choose techniques for collecting evidence of these slippery, abstract phenomena in the most appropriate way possible: as expressed by the perceivers themselves. To obtain the most reliable and valid results, it is important to design the methodology so as to avoid biases related to the researcher and the instruments. The researcher must encourage the user to focus on a real situation, relevant to the research questions, by using some kind of anchor for attention and recall. This is an advantage of time-line interviewing. The researcher must also be able to interpret the interview responses in a way that does not compromise the original expression of the user. This is an advantage of inductive content analysis.

Dervin and others suggest time-line interviewing as one means for providing a temporal framework to help users recall their cognitive states at certain times during their situations. Time-line interviewing has generally been directed toward information needs assessment and has been effective in studying users' perceptions in different situational contexts (e.g., Dervin, 1983, 1992; Fletcher, 1988; Gluck, 1993; Jacobson, 1991; Nilan, 1985; Nilan, Peek, & Snyder, 1988; Schamber, 1991).

Typically, the interview follows a series of open-ended items in a structured questionnaire administered by the researcher. Individual respondents are asked to describe their own situations as sequences of events. The events are noted on index cards that are laid out sequentially to form visual reference points for both respondent and researcher as they discuss various aspects of the situation. Additional cards are laid out beneath each event card for questions the respondent had at that time. For each question, respondents are asked about any answer they received, including whether or how the answer helped them.

Content analysis is a well-established set of techniques for making inferences from text about sources, content, or receivers of information. It is widely used in the social sciences for analyzing texts of all kinds, from open-ended responses to survey questionnaires to mass media newspapers, television, and books. Beyond its purely analytical purposes, content analysis serves observational purposes in identifying data in texts.

The analytic process requires the use of a coding scheme, which consists of categories and operational definitions for specific variables (e.g., images of a certain societal group). Content-bearing units are identified in the texts and coded for appropriate categories. Categories can be derived inductively from the texts being analyzed, adapted from previous studies, or adopted unchanged from previous studies. Inductive content analysis is particularly appropriate for research that takes a grounded theory approach, or which derives theory from data rather than verifies existing theory. The development of new schemes entails decisions about units of analysis, category construction, and coding procedures. Schemes are tested for clarity and consistency based on intercoder agreement or reliability ratios. Results are reported in qualitative and/or quantitative terms (Krippendorff, 1980).

The sample study below illustrates use of these techniques at a greater level of detail than is typically offered in research reports, and is followed by a discussion of their pros and cons.

## Sample Study

A variation of time-line interview techniques was developed for an exploratory study intended to identify and describe information users' criteria for their relevance judgments in a multimedia (multiple source) information environment. The study addressed two areas that information scientists had investigated only in limited fashion from the user's perspective: dimensions of the concept of relevance in users' evaluations of information retrieval effectiveness, and dimensions of relevance in multimedia environments. Of particular interest were noncontent criteria underlying users' perceptions of relevance, or criteria relating not to the quality of information per se, but rather to qualities of information sources and presentation formats that may (or may not) add value to information.

The research questions were:

1. What criteria do users mention when they evaluate the results of information searches in a multimedia environment?
2. How do users' criteria differ for the results of information searches using different types of media in a multimedia environment?

Relevance was defined as the user's perception of the quality of relationship between information and his/her information problem situation at a given time. This broad view, including the importance of users' situations, is described in work by Harter (1992), Schamber (1994), Taylor (1986), and many others. Relevance judgments were seen as users' decisions to accept or reject information based on the extent of its relevance to the situation. Relevance criteria were defined as values or qualities underlying users' judgments of relevance. Criteria could relate to any aspect of information or situation, including information source and presentation format. A multimedia environment was defined as an information environment offering different types of sources and presentation formats.

In this study, the situational context was information seeking about the weather. The subject domain of weather was chosen primarily because weather information is available through a variety of formal and informal sources and in many presentation formats. The respondents were occupational users of weather information, people who needed weather information in order to make decisions or perform tasks. There were 30 respondents: 10 each in construction, electric power utilities, and aviation. Occupational users were expected to be highly motivated and specific in their needs for weather information and weather information systems.

The sense-making model as described by Dervin provided a methodological as well as a conceptual foundation for the study. Data collection was through an adaptation of Dervin's techniques for open-ended time-line interviewing. The interviews were audiotaped, the tapes transcribed, and the transcripts subjected to inductive content analysis in order to identify and categorize criteria.

At least three assumptions affected the methodologies. The first was that relevance criteria are defined in the context of users' perceptions of their own information problem situations. This meant that criteria could only be elicited directly from users describing their own situations. The second assumption, based on previous studies, was that users share understandings at some level about common relevance criterion concepts and situational phenomena (such as information sources). In fact, respondents were expected to mention and thus validate some criteria that had been suggested in previous studies. In a practical sense, without shared understandings, it would be impossible to create a meaningful coding scheme. The third assumption was that users' self-reports are valid indicators of their perceptions about their situations. This did not mean that users' reports encompassed every perception they had. But the results of previous sense-making studies demonstrated that individuals discussing their own recent situations were able to recall and articulate a great deal of detail.

The following sections provide an overview of the research techniques. For a detailed description of the methodology and results, along with full text of the interview questionnaire and content analytic coding schemes, and examples of raw data, see Schamber (1991).

## Time-Line Interviews

Data were collected through structured interviews in which a time-line was used to establish sequential pictures of individual situations. For key events in their situations, respondents were asked to discuss the questions asked, answers received, sources consulted, and the presentation of information by sources. Interviews were conducted by the researcher in person in order to elicit the necessary detail in open-ended responses. The researcher was able to provide clarification for respondents and probe for further detail when necessary. In addition, the time-line interview technique required a face-to-face setting for the use of index cards as visual cues.

### *Questionnaire Development*

The process of developing the questionnaire included pretest interviews with 13 respondents. The questionnaire was revised after each interview. The goal, as in Dervin's approach, was to use the time-line to establish a situation-oriented frame of reference and to provide items that would encourage respondents to express themselves as freely and naturally as possible. Questionnaire items were neutrally worded to avoid using biasing or suggestive terms for relevance or relevance criteria. Permission statements and nonthreatening probes were used to encourage respondents to recall and elaborate on details.

The most conspicuous modification to Dervin's technique was in the use of the time-line after the entire sequence of events was laid out. From that point on, the time-line was designed to focus respondents' attention on just three events in their situations during which they most actively sought information about the weather. From weather-seeking events, the focus was then narrowed further to explore weather questions, weather information sources, and presentation of information by the sources.

The primary reason for narrowing the focus to three events was time constraints. The interviews were limited to about 90 minutes because this seemed to be most feasible for respondents being interviewed at work. During this period the interviewer could not possibly pursue all the events and questions in a situation to the depth required to collect open-ended evaluations of sources and presentations. This conclusion was reached during the pretest interviews when it became apparent that individuals in weather-related situations consulted far more sources, and tended to describe their criteria in more detail, than anticipated.

### *Questionnaire Structure and Administration*

The interview questionnaire consisted of 22 items in six sections: (A) Introduction, (B) Time-Line, (C) Question Loop, (D) Source Loop, (E) Idealization, and (F) Conclusion (see Table 1). Loops were series of items that were repeated for different questions users had or sources they consulted. Many items called for a closed response followed by an open-ended explanation or description. The questionnaire collected demographic, situational, and evaluational data. The sections that collected situational and evaluational data are described briefly below, along with operational definitions of key concepts.

The (B) Time-Line allowed data to be collected about an individual's situation at the levels of events, questions, and sources. Respondents were asked to talk about one recent job-related situation in which they required information about the weather in order to make a decision or perform a task. First they described the sequences of events in their situations. Events were operationalized as something respondents did (e.g., consult a source), thought, or felt at a certain time. Respondents then named the one event during which they were most active in seeking information about the weather (Weather Information Event). For this event and the event before and after it (Weather Event Set), they described all the questions (weather and nonweather) they had. A question was operationalized as something the respondent wanted to find out, understand, or make sense of; or concerned who to ask or where to go for an answer; or simply expressed a feeling. Finally, for each weather question, respondents named the sources they consulted for answers. Again, the narrowing of the time-line to pursue just three events was the major change from the way the technique was used in previous studies. Another change was focusing on sources for only certain types of questions.

The (C) Question Loop allowed respondents to evaluate the information they received with regard to their situations. When they were asked whether information partially or completely answered their questions, they were in essence making relevance judgments. The items concerning how answers did or did not help them were intended, in sense-making terms, to get at qualities of gap-bridging, or uses of information in resolving information problem situations. Although Dervin and others have relied heavily on responses to these items, they were not analyzed for this study because the emphasis was on source evaluation, not relevance judgments per se. In this study (which was part of a

TABLE 1. Interview questionnaire structure.

| | | Data type[a] | | |
|---|---|---|---|---|
| Item set | Response type | Demographic | Situation | Evaluation |
| A. Introduction | | | | |
| Age (birthdate on consent form) | open | * | | |
| B. Time-line | | | | |
| 1. Situation | open | | * | |
| 2. Events | open | | * | |
| 3. Critical event (for task) | open | | — | |
| 4. Weather event (information seeking) | open | | * | |
| 5. Weather event set (3 events) | open | | * | |
| 6. Weather event set questions (all) | open | | * | |
| 7. Sources (for weather questions) | open | | * | |
| C. Question loop | | | | |
| 8. Expected helps (in getting answer) | open | | | — |
| 9. Importance (of getting answer) | 0–6, open | | | — |
| 10. Got answer | n/y, open | | | — |
| 11. Tried to get answer | n/y, open | | | — |
| 12. Difficulty (of getting answer) | 0–6, open | | | — |
| 13. Answer (completeness) | partial/complete | | | — |
| 14. Actual helps (in getting answer) | open | | | — |
| D. Source loop | | | | |
| 15. Source made a difference | n/y, open | | | ** |
| 16. Presentation made a difference | n/y, open | | | ** |
| 17. Clarity (of answer) | 0–6, open | | | ** |
| 18. Change information (in answer) | n/y, open | | | — |
| E. Idealization | | | | |
| 19. Idealization (ideal presentation) | open | | | ** |
| F. Conclusion | | | | |
| 20. Experience in field | open | * | | |
| 21. Meteorology education | open | * | | |
| 22. General education | open | * | | |

[a] *Data were reported for this study; **data were content-analyzed for criteria; — data were not reported for this study.

larger study), the Question Loop served primarily to start respondents thinking evaluatively about their situations.

The (D) Source Loop collected the core data, criteria. Respondents were asked, first, whether a particular source made a difference to them in their situations. After they responded yes or no, they were asked to explain their responses further; to describe how or why the source did or did not make a difference. Second, they were asked whether the presentation of information by the source made a difference and again, to explain their yes or no response. The words "made a difference" operationalized the concept of a criterion in a neutral fashion. A third item asked respondents to rate the clarity of presentation by a particular source, then to explain their rating. The presentation and clarity items were added specifically for purposes of this study. The final item asked whether they changed the information in any way, and again to explain their response.

The (E) Idealization consisted of one item that was asked only once at the end of the interview. It asked respondents to describe how, ideally, information could have been presented in their situations. The purpose of the item was to encourage them to summarize and expand on their perceptions of the best possible combinations of presentation qualities, beyond the constraints of systems they used or knew about. The Idealization was also a new item added for this study.

Responses were recorded three ways. For the time-line, events and questions were written on index cards which, when laid out in order, provided a tangible set of reference points for orienting the discussion to the respondent's situation. Responses were also written in blanks on the questionnaire itself, with pages being added as necessary to iterations of the Question Loop and the Source Loop. The index cards and questionnaires served as backups for audiotapes of the interviews. The audiotapes were transcribed and the transcriptions used as primary sources of data for content analysis.

## Content Analysis

Content analysis served both as a secondary observation tool for identifying variables in interview texts, and as an analytic tool for assigning variables to categories in coding. The exploratory and descriptive purposes of this study required that coding schemes be developed inductively from natural-language data in the interview texts. Not only was this necessary for answering Research Question 1 about which criteria respondents mentioned, but it also applied to other variables such as questions, sources, and presentations.

### *Coding Scheme Development*

The development of all coding schemes involved the same general steps (Weber, 1985):

1. Determine the recording unit (word, phrase, sentence).
2. Develop categories.
3. Code data sample.
4. Test for intercoder agreement.
5. Revise and retest.

The criterion coding scheme was the result of a particularly long process of development, testing, and refinement. The process began during the pretest stage of developing the questionnaire, when the researcher informally identified criteria in interview responses in order to assess the effectiveness of the questionnaire items. The formal process of development began shortly after the first actual interviews. Based on data from the first few respondents, the scheme was significantly revised eight times and tested by some 14 coders until intercoder agreement reached acceptable levels (see below).

### *Coding Scheme Structure*

Eight coding schemes were derived from the interview texts:

1. Employment
2. Weather Question Type
3. Source Type
4. Presentation Type
5. Criterion Type
6. Criterion Focus
7. Criterion Presence
8. Criterion Desired Presence

Development and coding of the first four schemes was straightforward because categories were based on manifest content, or keywords, that were easy to identify. The (1) Employment scheme, which described job responsibilities, was demographic and the next three schemes were situational. The (2) Weather Question Type scheme categorized weather conditions (e.g., precipitation) about which respondents wanted information.

The (3) Source Type and (4) Presentation Type schemes were complicated somewhat by levels of mediation. For example, almost all weather information in the United States originates from one ultimate source, the National Weather Service. But it is usually filtered through mass media wire services, individual weathercasters, and dedicated information services (e.g., for aviation) before being disseminated to end-users. In addition, nonmediated weather information is available through weather instruments (e.g., thermometer, radar) and direct observation of actual weather conditions. However, direct observation—especially by airline pilots—also feeds into the National Weather Service. In this study, categories for both sources and presentations were defined at the level closest to the user as described by the user. For example, if the respondent reported seeing storm clouds, the source was coded Self and the presentation Direct Observation. If the respondent talked about television news maps, the source was Television and the presentation was Graphics (not text, newscaster, etc.). Each scheme had seven categories (plus Other). For Source Type they were Self, Other Person, System, Television, Radio, Newspaper, and Weather Instrument; for Presentation Type, they were Direct Observation, Interpersonal, Audio, Text, Graphics, Instrument Display, and Multimedia (combination).

The last four schemes presented challenges in both development and coding because they were based in part on latent, or contextual, content in natural language responses. The (5) Criterion Type scheme, which defined criteria, was central to the study, serving in itself as an answer to Research Question 1 on what criteria users mentioned. The intent was to describe all the criteria mentioned by respondents in respondents' own words. This scheme was repeatedly tested and revised. It contains 10 summary and 32 detail categories (Table 2). Three additional criterion schemes were also developed: (6) Criterion Focus, which identified whether a criterion referred to information, source, or presentation; (7) Criterion Presence, which identified whether a criterion quality was present or absent (e.g., accurate vs. inaccurate); and (8) Criterion Desired Presence; which identified whether a criterion quality was or was not needed or liked by the respondent.

The project Codebook (see Schamber, 1991) consisted of two parts: Data Coding and Entry, which explained procedures for handling all types of data (including quantitative), and eight Coding Schemes, which contained detailed category definitions, coding rules, and examples.

An example of category 40 in the Criterion Type scheme reads:

40 Reliability

Definition: Respondent trusted, believed, relied on, or had confidence in source and information from source; source was reputable.

Keywords: rely, reliable, credible, trust, trustworthy, reputation, had confidence/faith in, took stock in; unreliable, can't believe, untrustworthy, not credible.

TABLE 2. All levels of criteria mentioned in source/presentation responses[a]

| | Criterion mentions | | | | | |
|---|---|---|---|---|---|---|
| | Summary level | | | Both levels | | |
| Criterion | Freq. | Percent | Resp. | Freq. | Percent | Resp. |
| Accuracy | 43 | 5.3 | 20 | 43 | 5.3 | 20 |
| Currency | 114 | 14.1 | 27 | 52 | 6.4 | 19 |
| Time frame | | | | 62 | 7.6 | 23 |
| Specificity | 84 | 10.4 | 25 | 44 | 5.4 | 18 |
| Summary/interpretation | | | | 19 | 2.3 | 14 |
| Variety/volume | | | | 21 | 2.6 | 10 |
| Geographic proximity | 96 | 11.8 | 27 | 96 | 11.8 | 27 |
| Reliability | 107 | 13.2 | 26 | 48 | 5.9 | 22 |
| Expertise | | | | 34 | 4.2 | 17 |
| Directly observed | | | | 13 | 1.6 | 7 |
| Source confidence | | | | 4 | 0.5 | 1 |
| Consistency | | | | 8 | 1.0 | 5 |
| Accessibility | 52 | 6.4 | 20 | 4 | 0.5 | 4 |
| Availability | | | | 38 | 4.7 | 18 |
| Usability | | | | 8 | 1.0 | 6 |
| Affordability | | | | 2 | 0.2 | 1 |
| Verifiability | 103 | 12.7 | 26 | 64 | 7.9 | 24 |
| Source agreement | | | | 39 | 4.8 | 18 |
| Clarity | 34 | 4.2 | 16 | 7 | 0.9 | 4 |
| Verbal clarity | | | | 19 | 2.3 | 12 |
| Visual clarity | | | | 8 | 1.0 | 7 |
| Dynamism | 63 | 7.8 | 20 | 1 | 0.1 | 1 |
| Interactivity | | | | 39 | 4.8 | 16 |
| Tracking/projection | | | | 21 | 2.6 | 12 |
| Zooming | | | | 2 | 0.2 | 2 |
| Presentation quality | 115 | 14.2 | 25 | 5 | 0.6 | 5 |
| Human quality | | | | 31 | 3.8 | 15 |
| Nonweather information | | | | 5 | 0.6 | 4 |
| Permanence | | | | 13 | 1.6 | 10 |
| Presentation preference | | | | 34 | 4.2 | 12 |
| Entertainment value | | | | 6 | 0.7 | 3 |
| Choice of format | | | | 21 | 2.6 | 15 |
| Column total | 811 | 100.1* | — | 811 | 99.7* | — |
| Column mean | 81.1 | — | 23.2 | 25.3 | — | 11.6 |

[a] Based on responses to questionnaire items (15) Source Made a Difference and (16) Presentation Made a Difference by 30 respondents. Summary-level data (center columns) include data for detail categories under them. "Freq." is the number of responses in which a criterion was mentioned. "Resp." is the number of respondents who mentioned a criterion at least once.

* Does not equal 100 due to rounding error.

Examples: "I put a lot of credence in it." "I don't take much stock in the news weather." "Some of the guys don't take any faith in some of these systems."

Note: Reliance or trust may also depend on presentation of information, as in: "They tend to believe me more when I've got a piece of paper in my hand."

### Criterion Coding

The task of coding criteria was challenging because so many criteria appeared in the interview texts, appeared repeatedly in responses to the same items, and appeared in both manifest and latent content. For coding purposes, a response was all the text generated by an individual in response to one questionnaire item. Each response contained at least one coding unit: a word or group of words that could be coded under one criterion category (or coded No Data, which was rare). A response had to be unitized, or all the coding units identified, before it could be coded. An example is this response evaluating a dedicated weather information system:

> Reputation. Accurate. Usually up-to-date. They provide us updates on the storm, what's anticipated. Sometimes it's on weather maps. They usually update for us at any given time if there is a change in a previous forecast, or we could request at any time for them to send it to us or send us a weather map. It's all sent by telecopier. It could be text or a map or a combination of both.

This response was unitized and coded as follows:

| Coding unit | Criterion category |
|---|---|
| (1) "Reputation" | Reliability |
| (2) "Accurate" | Accuracy |
| (3) "up-to-date"<br>"provide us updates"<br>"update for us" | Currency |
| (4) "we could request at any time for them to send it to us or send us a weather map" | Availability |
| (5) "Sometimes it's on weather maps"<br>"It could be text or a map or a combination of both" | Choice of format |

## Data and Results

In the sample study, time-line interviews and inductive content analysis both served as methods of observation. Together they yielded not only an enormous amount of data, but also extremely rich data that could be examined for contextual implications.

### *Data Reporting*

Demographic data revealed, among other things, that the occupational users of weather information were highly experienced, averaging 20 years experience at an average age of 40.

Situational data described a wide range of weather-related planning decisions: the protection of workers and materials during winter construction projects, the scheduling of electric power line maintenance and repairs, and the scheduling and routing of airplane flights. Respondents had to make decisions based on essentially unpredictable weather conditions within serious time constraints and safety restrictions. The fact that decisions ultimately had life-or-death significance led to the impression that these respondents were highly motivated. The time-line interviews were successful at managing the varied and complex descriptions of these situations. Respondents described 3 to 11 events in their situations, with a mean of 6. These were narrowed to just 3, the Weather Event Set. During these events alone, all 30 respondents said they consulted weather information sources 189 times, or more than 6 times per respondent on average. Each consulted 1 to 7 different types of sources and presentations, or a mean of nearly 3 types each.

Evaluational data were drawn from content analysis of responses to four questionnaire items: Source Made a Difference, Presentation Made a Difference, Clarity, and Idealization. The 30 respondents generated 365 responses in which content analysis identified 1,199 mentions of criteria. On average, each respondent generated 12 responses that contained 40 criteria, or 3 criteria per response. Criterion frequencies were reported, first, in terms of the number of responses in which a criterion was mentioned, which varied considerably from one respondent to the next; and, second, in terms of the number of respondents who mentioned a given criterion at least once, which could not exceed the total number of respondents, 30 (see Table 2). Frequencies were not taken as indicators of the relative importance of criteria (beyond the fact that respondents considered criteria worth mentioning at all) because respondents were only asked to make evaluations, not to rate or rank criteria in any way. In addition, the totals of summary categories that had more detail categories beneath them may be exaggerated. The data were not appropriate for statistical testing, although correlation matrices were computed to help visualize the results for exploratory purposes.

### *Criterion Results*

The primary criterion results were based on responses to just two items: Source Made a Difference and Presentation Made a Difference. In answer to Research Question 1 concerning the definitions of criteria, the study successfully identified and described a range of criteria in 10 summary categories and 22 detail categories. The summary categories were Accuracy, Currency, Specificity, Geographic Proximity, Reliability, Accessibility, Verifiability, Clarity, Dynamism, and Presentation Quality (Table 2).

The answer to Research Question 2 concerned differences among criteria mentioned in evaluations of seven types of sources and seven types of presentations (not including Other). The results showed that in general, most criteria were mentioned regardless of source and presentation type, and criteria tended to appear more often for sources and presentations that were evaluated more often. It is interesting to note frequent mentions of the criteria Human Quality (e.g., personality), which implies awareness of and often preference for the human dimension; and Directly Observed (seeing or experiencing weather conditions), which implies trust in and often a preference for firsthand observation. Both criteria are unlikely to occur in traditional studies of user responses to a single text-based information retrieval system.

Further coding was able to shed more light on the meanings and uses of criteria. Coding for Focus revealed that a greater percentage of criteria seemed to be directed toward evaluating information than either source or presentation, despite the fact that respondents were only asked to evaluate sources and presentations. It also appeared that individuals were not evaluating sources without also evaluating information, or evaluating presentation without also evaluating source and information. Thus the coding categories were cumulative: Source included Information, and Presentation included both. The results of coding criteria for Presence (e.g., accuracy is present and inaccuracy is not) and Desired Presence revealed that the majority of criteria mentioned were both present in the situations and desired present by respondents.

Data based on responses to two additional questionnaire items helped describe the role played by presentation in the multimedia context. Results for the item concerning (17) Clarity showed that although clarity was not a major issue for expert users of weather information, the expert usually had to spend some time learning to read or understand the information presented. The (19) Idealization item was particularly effective in encouraging respondents to summarize their uses of criteria with respect to their situations. Most described in detail the ideal sources and presentations for their situations. A few said they had no preference for specific sources and presentations, as in: "I don't care if it comes to the TV or over my fax machine every night, so long as I could rely on it."

*Contextual Findings and Implications*

Although contextuality made the criterion mentions difficult to code, it provided rich fodder for interpretation. Not only did the results confirm a priori relevance variables from previous research, but the multimedia context also elicited nontopical criteria, such as those under Presentation Quality, that are unlikely to be considered in traditional text-based information retrieval studies. The results suggested many directions for future research. One is to examine similarities in co-occurring criteria (e.g., reliability and accuracy) and tradeoffs in inversely co-occurring criteria (e.g., sacrificing accuracy for currency). Another direction is to pursue situation type as a predictor of criterion use, with the idea of determining the elements and extent of situation types at the point where they effect criterion use. For example, despite widely varied user occupations in the weather-related context, the criteria mentioned were remarkably consistent, whereas a few criteria mentioned in a text-based information retrieval context have differed noticeably (see Barry & Schamber, 1998).

Dynamic aspects of evaluation behavior clearly should be studied. The sample study but did not exploit the full potential of time-line interviewing because the research questions did not address changes over time. The results indicate, however, that time and space were vital considerations. Respondents, who often were moving themselves in physical space, consulted a variety of sources, and consulted the same sources repeatedly, in order to monitor and verify constantly changing and unpredictable conditions. They actively used criteria to choose the sources with the best qualities for their purposes as a certain time and place.

Perhaps most important are the implications for system design. The results support Taylor's (1986) notion of value-added processes in systems, including his ideas about the feasibility of translating user evaluations into system features. In this study, time-line interviewing grounded respondents in situational time and space while allowing them freedom to describe a broad range of source qualities. They were remarkably clear about the tangible system features they used, required, or wanted. The idealization item was especially useful in encouraging them to project their values beyond the constraints of familiar systems. Although the responses ran the gamut from no preference to descriptions of very sophisticated information systems, the majority of respondents seemed to want every feature they had heard about, and more, to help them perform their jobs. The following is one of the shorter responses:

> I would definitely say visual aid. And whether I know how to read weather maps or not, something that shows maps, with a written explanation, and also look at the jet stream. Motion, as far as what the weather is supposed to look like. I've seen it where they show you what's going to happen if they speed it up like they do on TV. That's pretty interesting. They show you what's going to happen or what has happened. Something like that would be good. It'd be nice if a voice did tell you what it was. That would be the optimum situation, I guess, either the voice or written. I guess the voice would be the best thing. We're lazy—if somebody tells us something, it's easier than reading it. Again, I like to see the picture. That's most important.

This was coded in six detail categories: Verbal Clarity (under Clarity); Tracking/Projection (under Dynamism); and Presentation Preference, Entertainment Value, and Choice of Format (under Presentation Quality).

Generally, respondents expressed the desire for consistency and interactivity in sources, the ability to control output and display, and multiple sources for verification of information. Their behavior in monitoring multiple sources implies that it is unrealistic to expect one system to meet all of a user's needs all of the time. Their behavior in monitoring multiple sources suggests that their ideal systems should be designed for scanning as much as for directed retrieval. It also suggests that it would be unrealistic to expect one system to meet all needs all of the time in the weather-related context.

## Methodology Discussion

All research methods have pros and cons that are seldom discussed in traditional research reports. Researchers often find it difficult to locate detailed instructions for development, administration, coding, and so forth. Dissertations are particularly good at providing methodological details and full text of data collection and analysis instruments. In this article, it is hoped that an in-depth discussion, based on hard-won experience, can provide helpful insights for researchers who are considering using time-line interviewing and inductive content analysis.

*Advantages and Disadvantages*

*Time-line interview.* The time-line interview proved to be a useful tool, as other researchers have found, for orienting respondents to their situations and facilitating recall. It had several advantages:

- Time-lines were a naturalistic and relatively unobtrusive means of collecting data about respondents' cognitive perceptions. Establishing rapport and getting respondents to talk was not a problem; in fact, it was often necessary to slow them down and get them back on track.
- The structured questionnaire was a flexible instrument that allowed discussion of any number of elements—events, questions, and sources—in a respondent's situation, as well as in-depth discussion of specific elements.
- Recall did not appear to be a problem for respondents describing their own recent experiences. Recall was facilitated by the use of index cards as visual cues for events and questions and by nonthreatening probes that encouraged respondents to elaborate on their responses.
- The open-ended and neutrally worded items yielded rich data for content analysis.

The disadvantages concerned the labor-intensiveness of the methodology:

- The questionnaire was complex and difficult to develop. In order to serve as an effective instrument for both this study and a larger user study, it had to collect a great deal of data in a limited amount of time. The exact wording of the open-ended items was crucial to enhancing recall without introducing bias. Even after the questionnaire was refined through interviews with 13 pretest respondents and administered to 30 actual respondents, it was evident that more refinements could have been made.
- The questionnaire was complex and difficult to administer. It required training and experience on the part of the researcher, who under pressure of time had to work constantly to keep respondents on track and know when and how to probe for more depth and detail.
- The process of transcribing interview audiotapes and separating the transcript texts into codable response units for content analysis was time-consuming.

*Content analysis.* The inductive approach to content analysis was quite successful in identifying and defining relevance criteria as well as aspects of occupations, situations, sources, and presentations. It had several advantages:

- It was an unobtrusive adjunct to the interview method in that it was only performed after interviews, on the interview texts, and thus did not force theoretically defined concepts on respondents. In other words, the use of open-ended interview items plus subsequent content analysis lessened the potential for one kind of interview bias in that respondents were asked only to focus on their perceptions of their situations, not on generating "acceptable" responses framed in the language of the researcher.
- It served the exploratory and descriptive goals of the study. The inductive process of developing content analytic coding schemes not only refined the operational definitions of variables, but it also identified the ranges of variables. The Criterion Type coding scheme in itself answered Research Question 1 by describing the range and meanings of criteria.
- It was a necessary method for analysis of unstructured interview texts. It worked well with texts of different lengths and was sensitive to context in those texts.

The disadvantages, as with questionnaire development and interviewing, concerned the labor-intensiveness of the method, especially with regard to the Criterion Type scheme:

- The criterion coding scheme was complex and difficult to develop. The process involved identifying both manifest and latent meanings in natural-language text, then creating distinct categories for coding those meanings.
- The criterion coding scheme was complex and difficult to apply. The process of coding required training and experience on the part of coders, who had to understand the definitions of a large number of categories as well as be able to interpret apparent meanings contained in natural language text.
- Content analysis for this exploratory study did not yield data suitable for statistical analysis.

### *Reliability and Validity*

In discussing any research design at any level, it is important to account for reliability and validity, which determine the extent of conclusions that can be drawn. Because the sample study was intended to be exploratory and descriptive—not predictive or generalizable—it is also interesting to note how reliability and validity were assessed.

*Reliability.* The fact that the interview questionnaire was modified and the content analytic coding schemes created specifically for this study brings into question the consistency or potential repeatability of results using these instruments. In this case, the questionnaire was pretested 13 times and the same researcher conducted all 30 interviews using the same structured format.

The content analytic schemes were tested for intercoder reliability based on simple percent agreement: the number of agreements between two independent coders divided by the number of possible agreements. The minimum standard for acceptability for most studies has been established as 90% and for exploratory studies as 80% (Krippendorff, 1980).* For this study, achieving intercoder reliability of 90% or better was not a problem except for one scheme: Criterion Type. After multiple tests, intercoder reliability for Criterion Type was 81.8% at the summary level and 77.3% at the detail level for an inexperienced (first-time) coder. This was considered extremely good for a 32-category scheme. Nevertheless, there was no way to assess the effects of any individual interviewer or coder biases or potential variations among future interviewers and coders. Because it was evident that training and experience of interviewers and coders was necessary for reliable results,

*Krippendorff calls simple percent agreement "deceptive" (p. 135) and suggests several alternative agreement coefficients that account for chance agreement. He also warns that percent agreement between 67% and 80% should be used "only for drawing highly tentative and cautious conclusions" (p. 147).

instructions and rules were provided with the questionnaire and coding schemes to explain the process for future researchers (see Schamber, 1991).

*Content validity.* In the sample study, the fact that the focus was on identifying and defining a full range of relevance criteria made content validity the central concern. A claim for high content validity was made on the basis that:

- Data were elicited directly from users describing their own situations. In many studies, claims for content validity are based on the judgments of experts who agree that items represent dimensions of real-world phenomena, or simply on the researcher's judgment that an item appears to measure a concept adequately. Here respondents themselves, in voluntarily generating criterion mentions in open-ended responses, served as firsthand judges of the realness of the concepts.
- The Criterion Type coding scheme (and others) included definitions, examples, and rules intended to define the concepts in considerable detail and improve consistency in coding. Although the researcher necessarily imposed structure on the scheme, the criteria were defined, insofar as possible, in the respondents' own language.
- Frequency and redundancy of criterion mentions by individual respondents implied that the concepts were clearly understood and consistently applied in individual situations. The data in table 2 show that criteria were mentioned often and repeatedly.
- Redundancy of criterion mentions across 30 independent respondents indicated coverage of a full range of criterion concepts across three occupational fields (aviation, electric power utilities, construction) in the weather-related context. In order to describe the range of any concept, data must be collected until no new instances of any category of that concept appear, or until redundancy is reached. The content analysis identified 32 categories of criteria. If mentions of new criterion categories are taken in the order respondents were interviewed, 7 respondents were sufficient to generate 20 categories, 10 respondents to generate 31 categories, and 18 respondents to generate all 32 categories. (The fact that 10 or fewer respondents can generate nearly a full range of cognitive perceptions has been observed in other studies; see, e.g., Barry, 1994; Fletcher, 1988). All except one criterion, Clarity, was mentioned by 20 to 27 (66.7% to 90%) respondents, and even Clarity was mentioned by 16 (53.3%) of the respondents.
- Several factors indicate shared understandings of criterion meanings. Frequency and redundancy of criterion mentions demonstrate agreement among diverse users in the weather-related situations. Intercoder reliability demonstrates agreement between respondents and coders, researcher and coders, and coders and other coders. Finally, the fact that criterion types identified in this study strongly overlap those identified in similar user studies and overlap a priori relevance factors suggested in the literature (see Barry & Schamber, 1998; Schamber, 1994) seems to indicate face validity across types of situational contexts.

*External validity.* No claim for the generalizability of the results to other user populations was made, nor was this a major concern of the study. Respondents were purposively, not randomly sampled, on the basis of their occupational decision-making responsibilities in the weather-related context. Nevertheless, the redundancy of criterion mentions within this study and the overlap of categories with those of other studies imply the potential for generalizability of at least summary-level criteria across types of situational contexts.

### *Future Directions*

Methodological lessons learned from the sample study may inform future researchers. An exploratory descriptive study such as this requires clear research questions and careful structuring of the interview questionnaire. The questionnaire and content analytic coding schemes must be developed simultaneously through extensive pretesting. In the interview process, probing carefully for detail in open-ended responses will result in more clarity and consistency in responses, which will simplify both development of coding categories and actual coding of natural-language texts. Several additional modifications can be made to the time-line interview to simplify and improve the process for various purposes.

The flexibility of time-line interviewing is a decided advantage. Over the years the interview structure has evolved, and questionnaire items have been refined and modified, to fulfill various research purposes. In the sample study, narrowing the focus to just three events and adding three items to elicit evaluations of presentations was both necessary and useful for stimulating recall. Without sacrificing the value of situational orientation through the time-line, researchers can direct and narrow the focus in other ways.

The labor-intensiveness of interviewing can be reduced several ways. Several items that were not analyzed for this study but were included for a larger study can be dropped. Interviews can be administered to several respondents simultaneously. For example, Gluck (1993) used self-administered questionnaires in a group setting facilitated by the researcher (albeit at the admitted cost of some descriptive detail in responses). Fewer interviews can be conducted: researchers who are attempting to elicit cognitive perceptions for purely exploratory purposes can expect reasonably representative results with as few as 10 respondents.

An added item for this study that was particularly successful was the Idealization. While it is tempting to use the Idealization as the focus of an instrument, it is unlikely to work well without some kind of contextual anchor such as a time-line. In this study, it was probably effective because the interviewee had been talking for some time about the situation and making evaluations.

Time constraints limited the pursuit of users' perceptions in the weather-related situations to just three events. However, in view of the apparent importance of time in the situations and the criteria (e.g., Currency, Dynamism), pursuing events through the full time-line would have helped to answer another question about how uses of criteria change over time. Also, although it was not practical in this study to conduct interviews in real time (e.g., to observe pilots making crucial in-flight decisions), future researchers should consider doing so in order to elicit responses under the best possible conditions for recall and accuracy.

Researchers are often advised to replicate exploratory studies in an attempt to validate and improve the generalizability of results. Certainly users' relevance criteria should be—and have been—studied in other situational contexts. In this case, exact replication is not necessary using exactly the same labor-intensive methods. Studies by this researcher and others, using a variety of methods, have been able to validate the criteria using both qualitative and quantitative analytic techniques (see Schamber & Bateman 1999).

## Conclusions

Rich data is often cited as the reward of qualitative research, but seldom fully appreciated by researchers who have not had the opportunity to deal with it intellectually and methodologically through actual exploration and analysis.

The methods described here, time-line interviewing and inductive content analysis, were highly successful in eliciting, identifying, and defining relevance criteria as strongly grounded in real-life information problem situations. In fact, it is hard to imagine the value of discussing criteria outside some situational context. The results supported the assumptions that (1) criteria are best defined in the context of users' perceptions of their own situations, (2) users share understandings at some level about common criterion concepts and environmental phenomena, and (3) users' self-reports are valid indicators of their perceptions about their situations.

The flexibility of the methods makes them particularly suitable for exploratory work. The time-line can be readily adapted to focus on any situational area of interest. Of the three interview items added in this study to elicit evaluations of presentations, the idealization was especially effective in getting respondents to talk about source and presentation features beyond their current experience.

These are not research methods for the faint-hearted, however; they require time, effort, and skill in both development and use. A remarkable amount of control was required to avoid suggesting criterion concepts to respondents in the interviews and to overcome the classic challenges of interpreting natural language in content analysis. Considering the labor-intensiveness of the interviews, it is interesting to note that as few as 10 respondents can serve the purpose of eliciting cognitive perceptions such as relevance criteria.

On the whole, these methods are highly recommended for exploratory studies so long as researchers bear in mind—as with all methods—certain limitations.

## References

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45(3), 149–159.

Barry, C.L. & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. Information Processing & Management, 34(2/3), 219–236.

Dervin, B. (1983). An overview of sense-making research: Concepts, methods, and results to date. Paper presented to the International Communication Association, Dallas, TX.

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. (1997). Information seeking in context. In P. Vakkari, R. Savolainen, & B. Dervin (Eds.), Proceedings of the International Conference on Research in Information Needs, Seeking and Use in Different Contexts (pp. 13–38). London: Taylor Graham.

Fletcher, P.T. (1988). An exploration of situational dimensions in the information behaviors of general managers in state government. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Gluck, M. (1993). Understanding performance in information systems: An investigation of system and user views of geographic information. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Harter, S.P. (1992). Psychological relevance and information science. Journal of the American Society for Information Science, 43(9), 602–615.

Jacobson, T.L. (1991). Sense-making in a database environment. Information Processing & Management, 27(6), 447–457.

Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Beverly Hills, CA: Sage.

Nilan, M.S. (1985). Structural constraints and situational information seeking: A test of two predictors in a sense-making context. Unpublished doctoral dissertation, University of Washington, Seattle, WA.

Nilan, M.S., Peek, R.P., & Snyder, H.W. (1988). A methodology for tapping user evaluation behaviors: An exploration of users' strategy, source and information evaluating. Proceedings of the 51st Annual Meeting of the American Society for Information Science, 25, 152–159.

Schamber, L. (1991). Users' criteria for evaluation in a multimedia information seeking and use situations. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3–48.

Schamber, L., & Bateman, J. (1999). Relevance criteria uses and importance: Progress in development of a measurement scale. Proceedings of the 62nd Annual Meeting of the American Society for Information Science, 36, 381–389.

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

Weber, P.W. (1985). Basic content analysis. Beverly Hills, CA: Sage.