Chapter 4  Methodology

In this chapter, the initial section describes approaches to data collection methods for this study comparing the advantages and disadvantages of a natural setting and laboratory setting with respect to the research questions. It is followed by the section on sources of data decided to be collected. These include verbal protocols, search logs, and post-search interview. Then, the reasons of choosing scholars as sampled population is explained. The next two sections address the methods of data collection and data analysis. The last section discusses the justification of methodology taken in this study including reliability and validity issues.

4.1 Approaches to Methods

The first step needed to be done was to decide in which setting – laboratory setting or natural setting – the data should be collected. Tague-Sutcliffe (1992) provides a good comparison of these two settings. She explains that “A laboratory test is one in which the sources of variability stemming from users, databases, searchers, and search constraints are under the control of the experimenter” (p. 469). She also says that “By contrast, an operational test is one in which one or more existing systems – with their own users, databases, searchers, and search constraints – are evaluated or compared” (p. 469). She notes that there is a range from laboratory tests, with all four components (users, databases, searchers, and search constraints), to tests in which only one is controlled.

It is important to note that in this study, the difference between a natural setting and laboratory setting exists in controlling search constraints (e.g., tasks, time, and
physical place). Other factors mentioned by Tague-Sutcliffe make no difference to the setting to be used in this study. Users are going to be recruited from the same population (scholars), and the database that they will be searching on is the Web regardless of the setting in which the data is going to be collected.

There seems to be a consensus in which the determination of the level of control primarily depends on the research questions and the aims of a test. As Robertson (1981) illustrates, to answer a specific question, the research must be designed as a laboratory test to exclude any extraneous variations. On the other hand, according to Robertson, in order to answer a question that is directly related to real problems in the design of retrieval systems, and to provide answers which will apply to real situations, a test must be conducted in an operational environment. However, it is not always easy to characterize the research questions as either “specific” or “real situations.” Indeed, it is often the case that we are investigating a specific complex research problem which would apply to real situations.

Schamber (1994) notes that the trend has been toward operational settings, particularly for relevance-based evaluation. In fact, most recent studies on user-defined relevance criteria (Barry, 1994; Cool, Belkin, Frieder, & Kantor, 1993; T. K. Park, 1993; Schamber, 1991; Wang & Soergel, 1998) used the method of “natural conditions.” T. K. Park, in particular, emphasizes that the “naturalistic inquiry approach” is appropriate in understanding how users make selection decisions in accepting or rejecting information produced by IR systems.

The problem here is that the natural setting is not a perfect method for every study on users' evaluations in information systems because it has limitations in articulating
specific aspects of complex phenomena. For example, the naturalistic method collects
data after the evaluation is completed, not as it occurs. Therefore, it is difficult to
articulate the process of selection decision in accepting and rejecting. This is primarily
because people cannot remember everything that they have done or thought.

This study has taken a different approach from the previous studies on relevance
judgment. Instead of deciding a proper setting first, this study attempted to identify what
types of data need to be collected to answer each research question and a set of sub-
questions. The idea is that if we find out useful data collection method(s) for each
question, we could decide which setting (laboratory setting versus natural setting)
provides a better environment for collecting data utilizing particular method(s).

Eight possible data collection methods were identified: search logs, pre-search
interview, think-aloud, non-participant observation, post-search interview, diary, open-
end interview, and questionnaire. As seen in Table 3, for individual research questions
including sub-questions, the usefulness of data collection methods was marked as either
useful, somewhat useful, or not useful.
Research Question 1 (RQ1) deals with people’s decision-making process with respect to selecting an information object out of multiple choices. To answer RQ1, it is important and useful to collect concurrent verbal reports as the search proceeds. This is because people might have difficulties articulating decision processes after the searches are completed. Therefore, the “think-aloud” method seems to be the most useful data collection method to answer RQ1. Non-participant observation could provide some useful data, but has some limitations in answering “why” questions. Search logs also might not be able to provide useful data for answering “why” questions. They, however, could supplement post-search interviews as they help users recall their behaviors and judgments based on logs. It seems that post-search interviews (in a laboratory setting) could provide richer data than open-end interviews (in a natural setting). The diary and questionnaire methods probably could not be helpful in collecting data with regard to RQ1.

It appears that both post-search interviews (in a laboratory setting) and open-end interviews (in a natural setting) could provide useful data for Research Question 2 (RQ2), which deals with the extent of information quality and cognitive authority to people’s evaluation of information objects. Think-aloud would be a useful method for collecting verbal reports with regard to RQ2 as users express their evaluations and beliefs while they conduct searches in the Web. Other methods such as pre-search interview, non-participant observation, diary, and questionnaire do not seem to be useful or at most somewhat useful to answer sub-questions of RQ2.

The post-search interview would be the most useful method to collect data for Research Question 3 (RQ3). The only limitation to data collection through the post-
search interview is that it may be difficult to identify people’s “real” goals if tasks are
given to them. On the other hand, an open-ended interview could be useful for most of
questions of RQ3, but it is not as useful as the post-search interview. People’s think-
aloud protocols during the searches may not be so useful for RQ3. All other methods of
data collection including pre-search interview, non-participant observation, and
questionnaire may not be useful at all for RQ3.

In sum, Table 3 indicates that a laboratory setting provides a better environment
for data collection than in a natural setting because a greater number of questions can be
answered. One of the advantages of the laboratory setting is that it allows the researcher
to collect concurrent verbal reports through think-aloud methods, which provides
essential data for both Research Question 1 and Research Question 2. Also, post-search
interviews to be conducted in the laboratory setting could provide useful data for most
questions with regard to Research Question 3. Pre-search interview and non-participant
observation to be collected in the laboratory setting appear not to be useful for most of
the questions. Therefore, these two methods were dropped. Search logs are important
data to supplement post-search interviews in the natural setting. Questionnaires do not
seem to serve as a useful method for data collection for most of the research questions.
They can be used only as a supplementary instrument for collecting background
information about users.

4.2 Sources of Data

Data were collected from multiple sources, combining search logs, verbal
protocols during the searches, and interviews after the searches. The subjects were asked
to come to the laboratory to perform searches on the Web. Before they started searching,
they were asked to fill out a background questionnaire. During the searches, they were asked to "think aloud," their comments were recorded, and the searches were logged for subsequent use. After the subjects finished the searches, they were interviewed about their experiences during the searches. The principal advantage of this methodological approach is that it provides an opportunity to collect both concurrent and retrospective verbal data to observe users’ judgment of information. Also, utilizing search logs during the interview helped the subjects recall their judgment process vividly.

4.2.1 Verbal Protocols

Verbal protocol analysis is a research method that is widely used in cognitive science in order to understand what people are thinking while they are engaged in some tasks (Ericsson and Simon, 1993). This method has been used in a number of studies which investigated search behaviors in a full-text database (Nahl & Tenopir, 1996), online catalogs (Connell, 1995; Kiestra & Stockmans, & Kamphuis, 1994), and a ranked retrieval system (Keen, 1994; Belkin et al., 1998). These studies showed that subjects had little difficulty with explaining thoughts and reasons throughout the search sessions.

Think-aloud data is generally recognized as a major source of concurrent data on subjects’ cognitive processes in specific tasks. Without being constrained by a specific question, subjects verbalize their thoughts as they perform a task. This increases the amount of behavior that can be observed compared to the same subject working under silent conditions. Therefore, think-aloud data make it possible to relate subjects’ cognitive processes and experiences to observable behavior directly.
4.2.2 Search Logs

ScreenCam is a software program that captures PC screen activities and cursor movements, including clicking, scrolling, and typing in. Each search log with a task was saved as a separate file. The logs were collected for two purposes. One is to save logs for direct analyses in terms of Web pages that subjects looked at and actions that they took in the Web. The other is to utilize search logs during post-search interviews. The Lotus ScreenCam allows the experimenter to play back, pause, stop, and fast-forward the search logs during post-search interviews. The logs provide effective cues of subjects’ behaviors on a Web browser to both the experimenter and the subjects.

4.2.3 Post-Search Interview

The post-search interview was designed to elicit verbal reports by asking specific questions of subjects about their decisions and judgments during the searches. It was a semistructured, individual, face-to-face interview. Structured interviewing refers to a situation in which an interviewer asks each respondent a series of preestablished questions with a limited set of response categories while unstructured interviewing often refers to an open-ended ethnographic interview (Fontana & Frey, 1993). Semistructured interviewing stands between these two types of interview in which the interviewer directs the interaction and inquiry in a “somewhat” structured way.

While using a set of questions established (see Appendix B-10), there was a flexibility for variation in questions where the response may be probed. The interviewer controlled the pace of the interview, following interview questions in a standardized manner. However, the interviewer sometimes deviated from the sequence of questions
and question wording depending on the context or situation. Efforts were sometimes made to understand a particular term the subject used when it was directly related to the concept of quality and authority.

The terms “information quality” and “cognitive authority” were never used in the interviews. Instead, operationalized terms were used. For information quality, the questions included words such as “good,” “accurate,” “current,” and “correct,” and “useful.” Regarding cognitive authority, the questions included the word, “trust.”

The way interviews were conducted is described in detail in Section 4.4.3. Topic areas covered during the interviews are as follows:

- Reasons for selecting a particular site, search engine, or link
- Factors which influence decision-making with respect to their selection
- Evaluation of information object(s) with respect to goodness, currency, accuracy, usefulness, and trustfulness
- Characteristics of sources, information objects, and other situational factors that influence people’s judgment about information quality and authority.

4.3 Sample

Scholars were the sampled population in this study. Scholars were selected because they are more likely to be concerned with information quality and cognitive authority than other populations. A scholar’s work, by nature, is heavily involved in interaction with information: scholars find relevant information, assess the quality of the information, and use information in the research process. Therefore, most scholars feel competent to judge quality and authority based on their own evaluation criteria in the printed environment. Despite a long tradition of information use in the printed world,
scholars can be novice users in the Web since the Web is a relatively new information resource environment. The same rules and criteria they have used in the printed collection might not apply exactly to the Web environment.

Faculty members and doctoral students were included in the scholar sample. The reason for including doctoral students is as follows: doctoral students might have less experience in assessing information quality in the printed environment than faculty members, however, they might be as concerned with information quality for their research work as faculty members are. The subjects were also recruited from diverse disciplinary domains, as social scientists and natural scientists might have different ways of using the Web for their information problems.

Judgmental sampling (Krathwohl, 1993) was used to select a sample representative of the population. According to Krathwohl, in judgmental sampling, the researcher selects individuals “who are presumed to be typical of certain segments of the population and therefore representative of it” (p. 137). He says that this is a frequent strategy of qualitative research. Two main characteristics of the population in this study are: a) whether an individual is a doctoral student or a faculty member; and b) whether he/she is from social science area or natural science area. Therefore, the representatives of the population should be composed of at least four different individuals who are from different academic status and different discipline areas.

It was believed that scholars might be interested in participating in the experiment when they found that either methodology or topic of the study would be of interest to them. The following domain areas were identified as potentially useful: computer science, cognitive science, communication, and library and information science. The
flyer (Appendix A.1) was distributed through hard copies and email mailing lists to the
faculty members and doctoral students of the Department of Computer Science and
School of Communication, Information, and Library Studies in Rutgers University. The
recruitment letter (Appendix A-2) was sent out by email individually to doctoral students
and faculty members of the Department of Computer Science and School of
Communication, Information, and Library Studies. This letter was also sent out by an
e-mail to the faculty members of Rutgers Center for Cognitive Science in individual bases.
In addition, the researcher’s personal network was used to recruit subjects. As a result,
the 16 subjects, included doctoral students and faculty members from diverse discipline
areas, including computer science, chemistry, sociology, linguistics, computer
engineering, biomedical engineering, organizational psychology, communication, and
library and information science. Demographic characteristics of the subjects and their
experiences with the Web are described in Section 5.1.

4.4 Data Collection

4.4.1 Task

The subjects were given “generic tasks,” which outlined the kinds of task, but did
not restrict the specific information problems. The tasks were associated with research,
travel, medicine, and computer price information respectively. These four tasks were the
ones that were most frequently mentioned by the subjects in a previous study (Rieh &
Belkin, 1998). The dimensions of quality and authority were embedded in those tasks by
using such words as “good papers,” “useful information,” “credible information,” and
“best price.” Before each task is given to subjects, the following general written
instructions were given in order to provide the specification of the tasks:
Figure 3. General Instructions

Now you are going to be asked to do four different searches, for four different kinds of information. You will have up to 15 minutes to do each search. For each different search, you will be given a general "context," which you can specify to your own requirements and interests as required.

Please imagine that each of these four different contexts is actually relevant to you, and conduct your search as you normally would in such a situation.

The four tasks to be given in a written form are as follows:

Figure 4. Generic Tasks

1. For the research project in which you are currently engaged, you would like to find some good papers which are new to you, which you think will be useful.

2. You are planning for the next conference which you are going to attend, and would like to find useful information about hotels, restaurants, and features of interest in that city.

3. A friend of yours has just been diagnosed as having schistosomiasis, and you want to find credible information about the disease itself, and the best methods of treatment.

4. You've decided that you want to buy a new computer to use at home, and now you need to find the best price for it.

Tasks were given to the subjects one by one as they proceeded to search in a random order. Out of 24 possible arrangements with 4 different tasks, 10 cases were selected as a random sample using SPSS. When the 10 cases were used in order, from Subject 001 to 010, they were repeated from Subject 011 to 016. It was decided to assign tasks in the random order as opposed to be given in the same order to every subject because pilot tests found that the subjects mentioned searching to be tiresome by the time they reached
the third or fourth search. It was noted that the subjects’ being tired could affect their searching behaviors, and furthermore their judgment of information.

4.4.2 Experiment Setting

The experiments took place at the Information Interaction Laboratory at School of Communication, Information, and Library Studies in Rutgers University from December 1998 to June 1999. This Lab has facilities of video camera installed in the ceiling, a microphone, a PC, and a round desk where the subject can read a consent form, and fill out a questionnaire. The video camera focused on the monitor screen, and the face of subjects could not be recorded.

The PC used was equipped with a 100 MHz processor, 32 MB memory and 5.3 GB hard disk operating under Windows 95, using a 15” color monitor. Two kinds of Web browsers, Netscape 4.0 and Internet Explorer 4.0, were installed on the PC so that the subjects could choose the one with which they were more familiar. ScreenCam by Lotus Corp. was installed on the PC to save the search logs.

The experimenter was present in the laboratory during the searches to monitor the searches. She remained silent and unobtrusive during the searches except to prompt think-aloud utterances. The experimenter was seated behind the subject and hence was not visible to the subject. This arrangement made clear that social interaction between the subject and the experimenter was not intended when the subject did think-aloud (Ericsson and Simon, 1993).
4.4.3 Procedures

The procedures taken in the experiments are as follows:

On arrival, the subjects read and signed the Participant Information and Consent Form (Appendix B-1) explaining their activities, confidentiality of their responses, and disposition of audiotapes and videotapes.

They were asked to read the General Introduction to the Study (Appendix B-2) which described the scope of the study related to some questions in the Background Questionnaire.

They completed Background Questionnaire (Appendix B-3) which gathered background information and their experiences with the Web.

After completing the Questionnaire, they were ready to begin the searches. They received General Instructions of Tasks (Appendix B-4) which applied to the four tasks to be given. While they were reading the sheet, the experimenter went to the video-editing room, and started videotaping.

They were allotted 15 minutes to complete each search. They were instructed to “think-aloud” about what they were doing, and why they were doing it, as they searched. The videotape recorded the computer monitor during their searches and captured their "think-aloud" utterances.

They performed 4 searches based on generic tasks (Appendix B-5, B-6, B-7, B-8) that were given to them one by one. They searched on the Web using the Web browser they preferred. The experimenter took notes of questions or comments that the subject had (Appendix B-9). The entire search interaction was logged using Lotus ScreenCam.
After completing all four searches, the experimenter sat beside the subject to look at the computer monitor together. The experimenter opened the first search file, and clicked the “play” button of ScreenCam. For each movement of pages, at the exact moment of action (e.g., typing in words or URL, clicking a link or a button of a Web browser), the experimenter made the screen pause, and asked the subject the reasons for selecting a particular page, search engine, or a link. The subject answered the questions, and the experimenter made the screen continue to play. Once the screen displayed a new Web page, the experimenter asked a set of questions about the subject’s judgment about that page (see Appendix B-10 for interview questions). This was continued to the end of the search. This process was repeated four times with four different search files. The entire interview session was recorded both on videotapes and audiotapes.

4.5 Data Analysis

Out of 16 subjects, the data of one subject (S002) were dropped from the analyses, because this subject did not complete all four tasks. The audiotapes of these interviews for the remaining 15 subjects were transcribed, and the transcripts were marked by line number. The think-aloud utterances from 15 subjects were transcribed from the videotapes.

For the analysis of data, what needed to be done first was to integrate the data collected from three different sources: search logs, think-aloud data, post-search interview. That’s because it was necessary to analyze the subjects’ behaviors (from logs), concurrent verbal reports (from think-aloud), and retrospective verbal reports (from post-search interview) on one coding scheme form. Therefore, while listening to the interviews from audiotapes, search logs were played back using ScreenCam. It helped
the researcher identify a particular Web page or an action on which interview questions and answers were based. If an interview situation was not clear from a transcript, a videotape recorded during the interview session was played which showed the exact screen and utterances between the interviewer and the subject. The reason that videotapes were not used consistently for transcribing was that the resolution of the screen in the videotapes was not good enough to read text. However, it was good enough to show on which screen the question and answer was based.

The next step needed to be done was to develop a coding scheme through which the eventual categories were identified. The types of pages that subjects viewed, and types of action they took were categorized and coded from the search logs. The interview transcripts, merged with think-aloud utterances, were then subject to detailed analysis. The method of content analysis was then used as a technique to inductively identify and categorize judgments that the subjects made during the selection and decision-making processes, and factors and characteristics which influenced their judgments. Content analysis is a method “that uses a set of procedures to make valid inferences from text” (Weber, 1990, p. 9). According to Weber, a central issue in content analysis is that the many words of the text are classified into far fewer content categories. Each category may consist of one, several, or many words. Words, phrases, or other units of text classified in the same category are presumed to have similar meanings.

The basic unit of analysis was a Web page viewed by the subjects. Following Pharo’s (1999) definition, the term Web page is chosen to denote the “constellation of text, images, videos, and sounds that emerges as the result of a activating a URL in a
browser” (p. 208). For each page, a table was established to record the categories in Table 4. The example of data coding using this table is presented in Appendix C-4.

Table 4. Coding Form

|----|------|---------------------|-----------------------------------|--------------------|----------------------------------|----------------|

Following Ericsson and Simon (1993), the unit of analysis is the verbal statements made during searches and post-search interviews. According to Ericsson and Simon, “for the actual encoding, the protocol is segmented, and each segment corresponding to a statement” (p. 266). Statements could be a sentence, or were abbreviated to phrases or a single word.

The analysis of the responses indicated that the subjects were mentioning two distinct types of judgment. One type of judgment is predictive judgment, which is made before the subjects look at some Web page. The second type of judgment is evaluative judgment, made after they look at it. Consequently, the criteria for judgment were also classified as two types: criteria for predictive judgment and criteria for evaluative judgment. For easy discussion, the operational definitions of each category with examples from data are presented in Chapter 4. Here, the steps involved in the development of coding schemes are described as follows.

Identifying categories of evaluative judgment

Categories of evaluative judgment were embedded in the following interview questions such as “Do you believe that this information is good, accurate, current, or correct?”
“Do you think that this is useful information for your information problem?” When the subject mentioned the judgment of usefulness, goodness, trustworthiness, or other evaluation of pages during think-aloud, it was coded and categorized as evaluative judgment.

**Identifying categories of criteria for evaluative judgment**

Identifying categories of criteria for evaluative judgment was primarily based on the follow-up questions which were asked immediately after the questions about evaluative judgment. For example, when the subject responded to questions about usefulness, goodness, and trustworthiness, the next questions asked were: “If so, why do you think so? If not, why not?” or “If so, what makes you think so? If not, why not?”

**Identifying categories of predictive judgment**

Predictive judgment was embedded in interviewer’s questions which were asked when some movement was made by the subject from one page to another page. For example, related questions were: “Can you tell me why you started here?”; “What was it that made you to go to this site?”; “Why did you select this page to look at?”; “Why did you follow this link (what made you to follow this link)?”

**Identifying categories of criteria for predictive judgment**

Identifying categories of criteria for predictive judgment was primarily based on the follow-up questions which were asked immediately after the questions listed above regarding predictive judgment. For example, when the subject responded to questions about decision-making or selection, the questions were asked: “If so, why do you think so? If not, why not?”; “If so, what makes you think so? If not, why not?”
4.6 Justification of Methodology

The method used in this study is premised on the assumption that the users can identify and discuss the characteristics and features of information objects which influences on their judgments of information quality and authority. The previous studies (e.g., Barry, 1994; Cool, Belkin, Frieder, & Kantor, 1993; Park, 1993; Schamber, 1991; Wang & Soergel, 1998) demonstrated that users could in fact discuss the characteristics which influenced their relevance judgment process. These studies also consistently have shown that it was possible to summarize the kinds of criteria mentioned by developing a classification scheme which is based on the users’ own terms. This way of analyzing the data can be justified by a grounded theory which is a general methodology for developing theory. The theory is inductively grounded in data which are systematically gathered and analyzed (Strauss & Corbin, 1994).

To elicit the judgment of information quality and authority, the data were based on think-aloud utterances during the searches, and post-search interviews. Think-aloud during the searches means that the subject keeps on speaking out loud whatever thoughts come to mind while performing the task. There are no interruptions or questions as the subject is encouraged to give a concurrent account of his or her thoughts (Someren, Barnard, & Sandberg, 1994). This was a useful way to understand the subject's decision making and judgment during the actual cognitive processing.

However, because the thinking-aloud utterances are only a subset of all thoughts that occur while performing searches, and also because there are differences across users in articulation, the results can be viewed as indicative rather than conclusive (Wang & Soergel, 1998). To overcome this limitation, the interviews with the subjects were
conducted immediately after the searches were completed. It was noticed that the interview method was useful for several reasons: (a) It was able to collect responses which might not have been mentioned by the subject in think-aloud verbal protocols; (b) it would identify particular aspects of the subject's decision-making and judgment with respect to information quality and cognitive authority; (c) it would reveal detailed explanations of the subject's behaviors (e.g. entering queries, going to a Web site, selecting, not selecting, etc.).

In addition, the ScreenCam allowed us to play back the search logs including screen activities and cursor movements. This software helped the subjects remember the actions that they had taken during the searches. It also enabled us to pause and play the screen any time so that the subjects could answer the questions in length while interviewer was pausing the screen.

Having collected the data in these ways, it is still important to claim that the findings of this study are valid and reliable in the sense of representing real phenomena. Validity refers to the extent to which study findings are generalizable to a population (external validity) or to which the study accurately measures what it purports to measure (internal validity) (Heron, 1994). External validity is the power to create a consensus of generality beyond the circumstances in which it was studied (Krathwohl, 1993). Krathwohl describes several different kinds of evidence for validating the extent of generality: (a) subjects and situations, (b) treatment, (c) observation and measures, (d) time, (e) basis for sensing attributes or changes and procedure. As the sample of scholars was recruited from diverse backgrounds and academic status, they can be representative of scholars, but it is difficult to generalize the findings to a larger group of Web users.
The experiment settings were carefully designed, running 16 search sessions in the same laboratory with the same computer. Also, standardized data collection instruments were developed including general instructions (Appendix B-4), tasks in written forms (Appendix B-5, 6, 7, 8), and interview questions (Appendix B-10). All the experiments were conducted in the same laboratory. These kinds of standardized situations, equipment, and instruments increase the generality of this study.

Internal validity refers to whether the evidence of a study supports the existence of a relationship between or among its variables (Krathwohl, 1993). Krathwohl also defines it as the power to link variables in a relationship. The verification of internal validity in this study can be claimed by the fact that the data were collected through triangulation. By using information from two different data sources, think-aloud protocols and post-search interviews, it is possible to compare subjects' verbal reports in think-aloud utterances during the searches and interview responses after the searches. This methodology certainly has advantages for internal validity by providing better plausibility of the explanation and translation.

Reliability refers to the consistency of results produced (Krathwohl, 1993). In this study, the reliability issues involve the replicability of the data collection process and the consistency of content analysis. With regard to the data collection process, the concern is whether other researchers would have obtained the same results. Using search logs during the interview, rather than relying on the subject's memory, would provide a basis of the replicability of data collection. In addition, pre-structured interview questions would help the claim of replicability across different interviewers.

The consistency of content analysis was considered through intercoder agreement.
Two independent coders (Coder A and B) were recruited to code the set of 8 searches from two subjects. Each coder tested four searches from one subject on the bases of the guidelines outlined in Coding Definitions and Categories (Appendix C-2). According to Krippendorff (1980), “the lack of independence is likely to make data appear more reliable than they are” (p. 132). That is because discussions and communications among coders invariably influence coding toward a higher degree of agreement. Following his suggestions, the coders independently coded the interview transcripts and think-aloud protocols. The coding scheme of judgments (both predictive judgment and evaluative judgment) and criteria (both criteria for predictive judgment and criteria for evaluative judgment) were tested for reliability.

Inter-coder agreement was computed between the researcher and Coder A and B respectively. A widely used coefficient of reliability is the ratio of coding agreements to the total number of coding decisions (Holsti, 1969, p. 140):

\[
C.\ \text{R.} = \frac{2M}{N_1 + N_2}
\]

In this formula M is the number of coding decisions on which the two coders are in agreement, and \(N_1\) and \(N_2\) refer to the number of coding decisions made by judges 1 and 2, respectively.

Reliability for the coding of judgments was measured as follows. The researcher identified 91 judgments with the data from Subject 013, and Coder A identified 81 judgments from the same subject. We were in agreement on 68 of them. According to the formula, \(C.\ \text{R.} = \frac{2(68)}{91 + 81} = .79\). Another Coder B identified 93 judgments from the data of Subject 005 while the researcher identified 88 judgments. The number of items on which we agreed was 64. Therefore \(C.\ \text{R.} = \frac{2(64)}{93 + 88} = .70\).
Reliability for the coding of criteria was computed as follows. From the data of Subject 013, the researcher identified 76 criteria, and Coder A reported 61 criteria. We were in agreement on 48 of them. C. R. = $2(48)/76 + 61 = .70$. With respect to the data from S005, the researcher identified 78 criteria, and Coder B identified 82 criteria. The items agreed between us reached 57. Therefore, C. R. = $2(57)/78 + 82 = .71$.

In sum, reliability measures for the coding of judgment categories reached levels of .79 and .70 respectively between the researcher and two coders. Those for the coding of criteria categories reached levels of .70 and .71. Although the reliability measures obtained for coding categories are somewhat low, they are still at the “acceptable” level for drawing conclusions in qualitative studies (Krippendorff, 1980).

### 4.7 Summary

Three research questions were posed. First, how do people decide which information source(s) to look at when they make choices among multiple sources in the Web? Second, to what extent are people concerned with quality and authority when they search in the Web? Third, what are the characteristics and factors that influence people’s judgments about information quality and cognitive authority? The examination of these questions required a laboratory experiment environment for data collection in which both concurrent and retrospective verbal reports could be gathered through a think-aloud and post-search interview respectively.

Participants were 16 scholars from diverse discipline areas recruited at Rutgers University. Each subject performed four searches on four “generic” tasks given to them. The subjects were given 15 minutes to complete each search. The entire search interaction was logged. The subjects were instructed to “think aloud” about what they
were doing, and why, as they searched. After the subjects completed all four searches, a semi-structured interview was conducted about their search activities while search logs were playing back.

Out of 16 subjects, the data of one subject were dropped from the analyses, because this subject did not complete all four tasks. The tapes of the interviews and think-aloud tapes for the remaining 15 subjects were transcribed. Content analysis was used as a technique to inductively identify and categorize the type and facets of judgment and the criteria mentioned by the subjects. Taking three kinds of data (search logs, interview transcripts, and think-aloud transcripts) together, the coding scheme was developed. The basic unit of analysis was a Web page viewed by the subjects. For each page, the following items were coded: evaluative judgment, criteria for evaluative judgment, predictive judgment, and criteria for predictive judgment, and action on the page.

The internal validity of this study can be claimed by the fact that the data were collected through triangulation. Using three different kinds of data (search logs, think-aloud utterances, and post-search interviews) allowed the researcher to verify the data from one to another. Multiple data collection method also provides a basis for replicability of data collection process. The consistency of content analysis with respect to reliability was based on the tests of intercoder agreement. Reliability measures of judgment categories and criteria categories reached the acceptable levels.